概率图模型学习理论及其应用

赵 悦 著

清华大学出版社

清华大学出版社数字出版网站

WQBook 文泉书局

www.wqbook.com

ISBN 978-7-302-30206-3

9 787302 302063

定价：23.00元

概率图模型学习理论及其应用

赵 悦 著

清华大学出版社
北京

内 容 简 介

作为概率论和图论相结合的产物,概率图模型理论为解决智能信息处理领域中的不确定性、复杂性问题提供了直观而自然的方法。近年来它逐步成为计算机视觉、语音识别、数据发掘与知识发现等领域中一个十分重要的研究方向,在国际上的影响不断扩大。本书是系统论述概率图模型的基本理论、学习算法及其应用的中文专著,内容包括概率图模型基本概念;完整数据集的概率图模型的学习理论;不完整数据集的概率图模型学习理论;无向概率图模型学习;新型学习方法;概率图模型在计算机视觉、个人信用风险评估及语言识别领域中的应用等部分。本书从实例出发,由浅入深,直观与严谨相结合,并提供了详尽的参考文献。

本书的读者对象是相关专业的高年级本科生、研究生和科研人员。

图书在版编目(CIP)数据

概率图模型学习理论及其应用/赵悦著.--北京:清华大学出版社,2012.12
ISBN 978-7-302-30206-3

Ⅰ.①概…　Ⅱ.①赵…　Ⅲ.①概率论—图论—数学模型　Ⅳ.①O211 ②O157.5

中国版本图书馆 CIP 数据核字(2012)第 229699 号

责任编辑:闫红梅　赵晓宁
封面设计:傅瑞学
责任校对:焦丽丽
责任印制:李红英

出版发行:清华大学出版社
网　　址:http://www.tup.com.cn, http://www.wqbook.com
地　　址:北京清华大学学研大厦 A 座　　**邮　　编**:100084
社 总 机:010-62770175　　**邮　　购**:010-62786544
投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn
质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn
课 件 下 载:http://www.tup.com.cn,010-62795954
印 装 者:北京国马印刷厂
经　　销:全国新华书店
开　　本:185mm×260mm　　**印　张**:10.5　　**字　　数**:253 千字
版　　次:2012 年 12 月第 1 版　　**印　　次**:2012 年 12 月第1 次印刷
印　　数:1～2500
定　　价:23.00 元

产品编号:045127-01

前　言

最近十年间，概率图模型在数据挖掘和模式识别等诸多领域得到了广泛的应用，促进了其理论研究的迅速发展。概率图模型为系统地解决数据挖掘和模式识别问题提供了似然概率框架，这些问题可以看作是一类搜索问题，是从局部的、不确定的和不明确的观测变量中得到全局相关的结果。不像在一些其他研究领域中，研究者们可能不用完全了解问题内部变量间的相关性，但是在数据挖掘和模式识别任务中，研究者们必须清楚变量间的相关性才能够建立一套准确、可靠的物理模型以实现指导整个搜索过程。图模型理论为这些研究提供了一套统一的框架，这套框架能够表达出变量间的相互关系和现实世界某些特殊问题的上下文关联知识，同时这套模型也能够通过严格的推理进行目标识别和分类。概率图模型较容易地捕获观测变量间及观测变量同领域知识间的相关性，同时能够进行系统的量化计算和不确定性推理。因此，在计算机视觉、语音识别和信用风险评估等模式识别和数据挖掘的任务中，概率图模型成为常用的模型工具。

概率图模型的发展历史可以追溯到 20 世纪 80 年代后期，Judea Pearl 和 Steffen Lauritzen 首先把图模型概念引入到人工智能和统计学习领域。不久以后，图模型迅速应用到统计、系统工程、信息论、模式识别和计算机视觉领域。在这些领域的广泛应用促进了概率图模型理论的发展，各种类型的模型结构、学习和推理算法层出不穷。关于概率图模型有两个主要的问题——学习和推理。学习主要是从数据中估计模型的结构和参数，然而推理的任务主要是对于给定的模型计算边缘和条件分布。和这两个问题相关的是计算复杂度问题，通常精确推理和学习是 NP-hard 问题。随着概率图模型学习理论和推理算法的不断成熟，其能够通过对大量数据的理解来建立精确的模型并且提供快速和有效的推理算法，这些都使得其具有独一无二的表示复杂领域知识的能力，使其继续成功地应用在更加广泛的领域内，如医疗诊断、军事决策、商业智能等。

基于概率图模型理论在越来越多领域中发挥的重要作用，其模型的学习理论和方法对其应用的有效性发挥了重要的作用，因此本书主要针对概率图模型学习理论进行梳理与总结，并结合其在不同领域的应用，尤其是计算机视觉、信用风险评估和语音识别中的应用，介绍最新的概率图模型学习理论的发展。

本书是作者在从事十多年的概率图模型研究工作的基础上编写的。书中较全面地介绍了概率图模型的基础知识、主要学习理论与计算方法。作者在国内外学术刊物与国际学术会议上发表了近 200 多篇论文，书中介绍了作者在这些论文中报道的部分研究成果。

本书共分 9 章。第 1 章介绍概率图模型的相关基本概念，主要涉及概率论、图论和信息论等知识；第 2 章重点对概率图模型的分类、学习及其推理原理进行详细介绍；第 3 章重点阐述传统的基于完整数据集的概率图模型参数学习和结构学习理论；第 4 章阐述基于不完

整数据的概率图模型学习方法；第 5 章介绍无向概率图模型的学习方法；第 6 章详细介绍近几年提出的新型概率图模型的学习算法；第 7～9 章介绍概率图模型理论在计算机视觉、个人信用风险评估、语音识别等领域的应用及相关问题分析。

本书撰写过程中参考了大量的国内外相关书籍和资料，在此对作者们表示衷心的感谢。

由于编者才疏学浅，书中错误在所难免，敬请读者批评指正。编者的 E-mail 为 zhaoyueso@sina.com。

编　者

2012 年 11 月

目　　录

第1章 概率图模型基础知识

1.1 概述

概率图模型是图模型理论和概率论相结合的产物，它为利用数学和工程的方法来解决不确定性和复杂系统问题提供了自然的工具。作为建模和推理工具，概率图模型理论使用直观、有效、灵活的图结构来表达多个随机变量的概率分布。在图结构中，结点代表了随机变量，结点间的连接(边)代表了随机变量之间的统计关系。图结构刻画了随机变量间的条件依赖和独立关系，从系统的角度揭示变量间的不确定性，并为不确定性的传递提供手段。基于概率图模型，联合概率分布可以分解成多个随机变量子集的函数，这个因式分解极大地简化了多变量联合概率分布模型。基于概率论，概率图模型算法利用图结构有效的计算边缘概率或其他条件概率。

在讨论概率图模型的基本原理和学习方法之前，本章就图论、概率论和信息论的相关概念予以介绍。

1.2 图论的相关基本概念

为了理解概率图模型，需要先了解一些图论的基本知识。

1. 无向图

一个无向图U是一个二元组$<N,E>$，即$U=<N,E>$，其中：

- N是一个非空集合的顶点集，N中的元素称为顶点或结点。
- E是无序积$N\times N$的多重子集(元素可多次出现)，称E为U的边集，E中的元素称为无向边或简称边。

在一个图$U=<N,E>$中，为了表示N和E分别为U的顶点集和边集。常将N记成$N(U)$，而将E记成$E(U)$。

2. 有向图

一个有向图D是一个二元组$<N,E>$，即$D=<N,E>$，其中：

- N是同无向图一样的顶点集。
- E是卡氏积的多重子集，其元素称为有向边，也简称边或弧；有时用$N(D)$、$E(D)$分

别表示图 D 的顶点集和边集。

在实际应用中，不论是无向图还是有向图，一般只画出它的图形，而不写出 N 和 E 的集合表达式。下面分别列出无向图和有向图的图形表示和对应的集合表示。

无向图 $U=\langle N,E\rangle$ 如图 1.1 所示，有

$$N=\{n_1,n_2,n_3,n_4,n_5\}$$

$$E=\{(n_1,n_2),(n_2,n_2),(n_2,n_3),(n_1,n_3),(n_1,n_4)\}$$

有向图 $U=\langle N,E\rangle$ 如图 1.2 所示，有

$$N=\{n_1,n_2,n_3,n_4,n_5\}$$

$$E=\{\langle n_1,n_1\rangle,\langle n_2,n_3\rangle,\langle n_3,n_2\rangle,\langle n_3,n_4\rangle,\langle n_2,n_4\rangle,\langle n_4,n_5\rangle,\langle n_5,n_4\rangle,\langle n_1,n_2\rangle\}$$

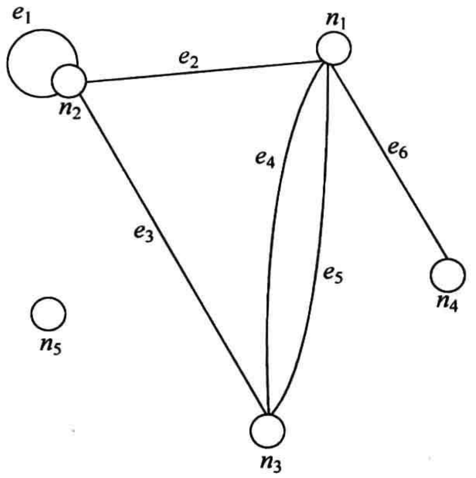


图 1.1 无向图

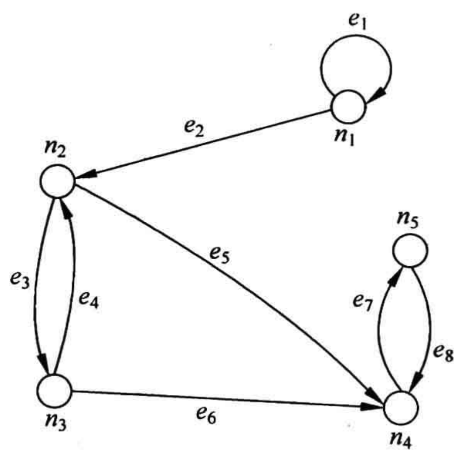


图 1.2 有向图

3. 混合图

如果在图中一些边是有向边，另一些边是无向边，则称这个图是混合图。

4. 邻接集(Adracency Set)

在一个图中，若两个结点由一条有向边或一条无向边关联，则称这两个结点互为邻接点。给定一个无向图 $U=\langle N,E\rangle$ 和图的一个结点 $n_i\in N$，则 n_i 的邻接集就是在图中直接和 n_i 相连的结点集合，即 $\mathrm{Adr}(n_i)=\{n_j\mid(n_i,n_j)\in E\}$。根据邻接点的概念，在有向图中也存在邻接集。根据有向边描述的方向性，在有向图中 n_i 的邻接集又可以分为两部分。邻接集的概念给出了图中结点间的邻域关系，用 $\Omega(n_i)$ 来表示结点 n_i 的邻域集合。

例如，在图 1.1 中，各个结点的邻接集为

$$\mathrm{Adr}(n_1)=\{n_2,n_3,n_4\};$$

$$\mathrm{Adr}(n_2)=\{n_1,n_2,n_3\};$$

$$\mathrm{Adr}(n_3)=\{n_1,n_2\};$$

$$\mathrm{Adr}(n_4)=\{n_1\};$$

$$\mathrm{Adr}(n_5)=\phi;$$

5. 有限图、m 阶图和平凡图

设 $G=<N,E>$ 为一个有向图或无向图，可定义以下概念：
(1) 有限图。若 N、E 都是有穷集合，则称 G 是有限图(在概率图模型中，只涉及有限图)。
(2) m 阶图。若 $|N|=m$($|N|$ 表示集合 N 的元素个数)，则称 G 为 m 阶图。
(3) 平凡图。若 $E=\phi$，则称 G 为零图。特别是，若此时又有 $|N|=1$，则称 G 为平凡图。

6. 平行边

在图中，连接同一对结点间的多条边称为平行边。含有平行边的任何一个图被称为多重图。

7. 简单图

不含有平行边和自环的图称为简单图(本书讨论的都是简单图)。

8. 无向完全图

设 $G=<N,E>$ 是 n 阶无向简单图，若 G 中任何顶点都与其余的 $n-1$ 个顶点相邻，则称 G 为 n 阶无向完全图。

9. 有向完全图

有向完全图 $G=<N,E>$ 设为 n 阶有向简单图，若对于任意的顶点 $u,v\in N,u\neq v$，既有有向边 $<u,v>$ 又有 $<v,u>$，则称 D 是 n 阶有向完全图。

10. 子图和完全子图

设 $G=<N,E>$，$G'=<N',E'>$ 是两个图，若 $N'\subseteq N$ 且 $E'\subseteq E$，则称 G' 是 G 的子图，G 是 G' 的母图。记做 $G'\subseteq G$。如果 G' 是完全图，则称 G' 是 G 的完全子图。

11. 极大完全子图与簇(Clique)或团

给定一个无向图 G，如果一个完全子图不是其他任何一个完全子图的真子图，则称为极大完全子图。极大完全子图的结点构成的集合又称为簇或团。

容易知道，完全图的任意子图都可能是完全图，只要所有连接子图结点的边都在子图内。一个图可能有很多子图。这其中应该会有一些“极大的”完全子图。

12. 通路和回路

给定图 $G=<N,E>$，设 G 中的顶点和边的交替序列为 $\Gamma=n_0e_1n_1e_2\cdots n_Le_L$，若 Γ 满足以下条件：n_{i-1} 和 n_i 是 e_i 的端点(在 G 是有向图时，要求 n_{i-1} 是 e_i 的始点，n_i 是 e_i 的终点)($i=1,2,\cdots,L$)，则称 Γ 为顶点 n_0 至 n_L 的通路。n_0 和 n_L 分别称为此通路的起点和终点，Γ 中边的数目 L 称为 Γ 的长度。当 $n_0=n_L$ 时，此通路称为回路。

13. 无环图和有环图

如果图中没有回路，则称为无环图(Unloopy Graph)。对应有回路的图称为有环图

(Loopy Graph)。

给定的概率图模型是否有环路,对信念传播算法非常重要。如果是无环图,信念传播算法可以得到准确的计算,其收敛性可以证明。如果是有环图,信念传播算法则大多采用一种近似算法。

14. 连通性和可达性

在一个无向图 U 中,若从顶点 n_i 到 n_j 存在通路(当然从 n_j 到 n_i 也存在通路),则称点 n_i 与 n_j 是连通的。规定点 n_j 到自身总是连通的。在一个有向图 D 中,若从顶点 n_i 到 n_j 存在通路,则称点 n_i 到 n_j 可达。规定点 n_i 到自身总是可达的。

15. 连通图和非连通图

若无向图 U 是平凡图,或 U 中任意两个顶点都是连通的,则称 U 是连通图;否则,称 U 是非连通图。

16. 连通分支

无向图中,顶点间的连通关系是等价关系。设 U 为一个无向图,R 是 U 中顶点之间的连通关系,按照 R 可将 $N(U)$ 划分成 $k(k \geq 1)$ 个等价类,记为 $N_1, N_2, \cdots, N_k$。由它们导出的子图称为 U 的连通分支,其个数记为 $P(U)$。

17. 点割集与割点

设无向图 $U=\langle N,E \rangle$,若存在顶点子集 $N' \subset N$,使 U 删除 N'(将 N' 中的顶点及关联的边都删除)后,所得 $U-N'$ 的连通分支数与 U 的连通分支数满足 $p(U-N')>p(U)$,而删除 N' 后的任何真子集 N'' 后,由 $p(U-N'')=p(U)$ 则称 N' 为 U 的点割集。若点割集中只有一个顶点 n,则称 n 为割点。

18. 相关性分割(D-Separation)

设 A、B、C 为有向图 D 中两两不相交的结点集,且 A、B 间的任意路径都被 C 阻塞,则称 A、B 被 C 相关性分割,而 C 称为 A、B 的切割集。

19. 父结点与子结点

在有向图中,连接同一条弧(边)的两个端点,根据弧的方向分为弧尾和弧头。有时也称弧尾结点是弧头结点的父结点,弧头结点是弧尾结点的子结点。

1.3 概率论的相关基本概念

本节主要介绍概率论中与概率图模型密切相关的一些基本概念。

1.3.1 随机变量与概率函数

1. 随机变量

如果对于试验的样本空间$\Omega=\{\omega\}$中的每一个样本点ω，变量X都有一个确定的实数值与之对应，则变量X是样本点ω的实函数，记做$X=X(\omega)$，称这样的变量X为随机变量。常用大写字母X、Y、Z等表示随机变量，用小写字母x、y、z等表示其取值。

随机变量的取值情况又可把其分为两类，即离散随机变量和连续随机变量。

2. 概率函数

设X为一随机变量，x是它的一个取值。在样本空间中，所有使X取值为x的原子事件组成一个事件，记做$\Omega_{X=x}=\{\omega\in\Omega \mid X(\omega)=x\}$，也记做“$X=x$”。事件“$X=x$”的概率$P(X=x)$依赖于$X$的取值$x$，让$x$在随机变量$X$的状态空间$\Omega_X$上变动，$P(X=x)$就成为$\Omega_X$的一个取值于$[0,1]$的函数，称之为随机变量$X$的概率质量函数，记做$P(X)$。

离散随机变量有概率质量函数与之对应，连续随机变量有概率密度函数与之对应。

1.3.2 古典概率与主观概率

在古典概率中，如果基本事件的总数为n，事件X所包含的基本事件个数为$r(r\leqslant n)$，则定义事件X的概率$P(X)$为r/n，即

$$P(X)=\frac{r}{n}=\frac{X\text{中包含的基本事件个数}}{\text{基本事件总数}} \tag{1.1}$$

主观概率又称为似然率，是人们对某一事件X发生信任程度大小的主观评价，即

$$P(X)=[\text{对}X\text{发生的信用度}] \tag{1.2}$$

1.3.3 联合概率分布

设(X,Y)是二维随机变量，对于任意实数x,y，事件$\{X\leqslant x\}$，$\{Y\leqslant y\}$同时发生的概率为

$$F(x,y)=P\{X=x,Y=y\} \tag{1.3}$$

称其为二维随机变量(X,Y)的分布函数，或称其为随机变量X和Y的联合分布函数。

1. 离散型随机变量的联合概率分布

如果二维随机变量(X,Y)只取有限个或可列个数对(x_i,y_j)，则称(X,Y)为二维离散型随机变量，称

$$P\{X=x_i,Y=y_j\}=p_{ij}\quad i=1,2,\cdots;\ j=1,2,\cdots \tag{1.4}$$

为(X,Y)的概率分布，或X与Y的联合概率分布。

2. 连续型随机变量的联合概率分布

如果存在二元非负函数$f(x,y)$，使得二维随机变量(X,Y)的联合分布函数$F(x,y)$可表示为

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(x,y)\mathrm{d}x\mathrm{d}y \tag{1.5}$$

则称(X,Y)为二维连续型随机变量，称$f(x,y)$为(X,Y)的概率密度，或X与Y的联合概率密度。

1.3.4 边缘概率分布

设二维随机变量(X,Y)具有分布函数$F(x,y)$。X和Y都是一维随机变量，也各有对应的分布函数$F_X(x)$和$F_Y(y)$，依次称为二维随机变量(X,Y)关于X和关于Y的边缘分布函数。

易知

$$F_X(x)=P\{X\leqslant x\}=P\{X\leqslant x,Y<+\infty\}=\lim_{y\to+\infty}F(x,y)=F(x,+\infty) \tag{1.6}$$

$$F_Y(x)=P\{Y\leqslant y\}=P\{X<+\infty,Y\leqslant y\}=\lim_{x\to+\infty}F(x,y)=F(+\infty,y) \tag{1.7}$$

1. 二维离散型随机变量的边缘分布

设二维离散型随机变量(X,Y)的联合分布为

$$P\{X=x_i,Y=y_j\}=p_{ij}\quad i=1,2,\cdots;\ j=1,2,\cdots$$

则称

$$p_i=P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij}\quad i=1,2,\cdots \tag{1.8}$$

为(X,Y)关于X的边缘分布。

称

$$p_j=P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij}\quad j=1,2,\cdots \tag{1.9}$$

为(X,Y)关于Y的边缘分布。

2. 二维连续型随机变量的边缘概率密度

设二维连续型随机变量(X,Y)的分布函数为$F(x,y)$，概率密度为$f(x,y)$。因为

$$F_X(x)=F(x,+\infty)=\int_{-\infty}^{x}\left(\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y\right)\mathrm{d}x \tag{1.10}$$

由分布函数定义知，X是一个连续型随机变量，且其概率密度为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y \tag{1.11}$$

同样有

$$F_Y(y)=F(+\infty,y)=\int_{-\infty}^{y}\left(\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}x\right)\mathrm{d}y \tag{1.12}$$

所以，Y也是一个连续型随机变量，其概率密度为

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}x \tag{1.13}$$

称$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y$为$(X,Y)$关于$X$的边缘概率密度；称$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)$

$\mathrm{d}x$ 为(X,Y)关于 Y 的边缘概率密度。

1.3.5 条件概率分布

对二维随机变量(X,Y)而言,随机变量 X 的条件概率分布,就是在给定 Y 取某个值的条件下 X 的概率分布。

1. 离散型随机变量的条件分布

设二维离散型随机变量(X,Y)的联合概率分布为

$$P\{X=x_i,Y=y_j\}=p_{ij} \quad i=1,2,\cdots;\ j=1,2,\cdots$$

对一切使 $P\{Y=y_j,\}=p_j=\sum\limits_{i=1}^{\infty}p_{ij}>0$ 的 Y_j,称

$$p_{i|j}=P\{X=x_i\mid Y=y_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_j} \quad i=1,2,\cdots \tag{1.14}$$

为给定 $Y=y_j$ 条件下 X 的条件概率分布。

两变量 X、Y 的联合分布 $P(X,Y)$,按照条件概率分布的定义,可得

$$P(X,Y)=P(X)P(Y\mid X) \tag{1.15}$$

将其推广到 n 个变量的联合分布 $P(X_1,X_2,\cdots,X_n)$,有

$$P(X_1,X_2,\cdots,X_n)=P(X_1)P(X_2\mid X_1)\cdots P(X_n\mid X_1,\cdots,X_{n-1}) \tag{1.16}$$

将一个联合概率分布分解为一系列条件分布的乘积,称为链规则。

2. 连续型随机变量的条件分布

设(X,Y)的概率密度为 $f(x,y)$,(X,Y)关于 Y 的边缘概率密度为 $f_Y(y)$,给定对于任意固定的 x,对于任意 $\varepsilon>0$,考虑条件概率 $P\{X\leqslant x|y<Y\leqslant y+\varepsilon\}$,设 $P\{y<Y\leqslant y+\varepsilon\}>0$,则有

$$P\{X\leqslant x\mid y<Y\leqslant y+\varepsilon\}=\frac{P\{X\leqslant x,y<Y\leqslant y+\varepsilon\}}{P\{y<Y\leqslant y+\varepsilon\}}=\frac{\int_{-\infty}^{x}\left[\int_{y}^{y+\varepsilon}f(x,y)\mathrm{d}y\right]\mathrm{d}x}{\int_{y}^{y+\varepsilon}f_Y(y)\mathrm{d}y} \tag{1.17}$$

在某些条件下,当 ε 很小时,式(1.17)右端分子、分母分别近似于 $\varepsilon\int_{-\infty}^{x}f(x,y)\mathrm{d}x$ 和 $\varepsilon f_Y(y)$,于是当 ε 很小时,有

$$P\{X\leqslant x\mid y<Y\leqslant y+\varepsilon\}\approx\frac{\varepsilon\int_{-\infty}^{x}f(x,y)\mathrm{d}x}{\varepsilon f_Y(y)}=\int_{-\infty}^{x}\frac{f(x,y)}{f_Y(y)}\mathrm{d}x \tag{1.18}$$

因此,设二维随机变量(X,Y)的概率密度为 $f(x,y)$,(X,Y)关于 X、关于 Y 的边缘概率密度分别为 $f_X(x,y)$和 $f_Y(x,y)$,若对于固定的 y,$f_Y(y)>0$,则称 $\dfrac{f(x,y)}{f_Y(y)}$为在 $Y=y$ 的条件下 X 的条件概率密度,记为

$$f_{X|Y}(x\mid y)=\frac{f(x,y)}{f_Y(y)} \tag{1.19}$$

若对于固定的 x，$f_X(x)>0$，则称 $\frac{f(x,y)}{f_X(x)}$ 为在 $X=x$ 的条件下 Y 的条件概率密度，记为

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} \tag{1.20}$$

1.3.6 边缘独立与条件独立

1. 边缘独立

设(X,Y)的联合分布为 $P(X,Y)$，边缘分布为 $P(X)$和 $P(Y)$，若式(1.21)成立，即

$$P(X,Y) = P(X) \cdot P(Y) \tag{1.21}$$

则称 X 与 Y 相互(边缘)独立。

2. 条件独立

随机变量 X、Y、Z，设 $P(Z=z)>0$，$\forall z\in\Omega_Z$，如果式(1.22)成立，即

$$P(X,Y,Z) = P(X \mid Z)P(Y \mid Z) \tag{1.22}$$

则称 X 和 Y 在给定 Z 时相互条件独立。

1.3.7 贝叶斯定理

贝叶斯定理又称为贝叶斯规则或贝叶斯公式。设 H 和 E 为两个随机变量，$H=h$ 为某一假设，$E=e$ 为一组证据。在考虑证据 $E=e$ 之前，对事件 $H=h$ 的概率估计 $P(H=h)$称为先验概率，而在考虑证据之后，对 $H=h$ 的概率估计 $P(H=h|E=e)$称为后验概率。贝叶斯定理描述了先验概率和后验概率之间的关系，即

$$P(H = h \mid E = e) = \frac{P(H = h)P(E = e \mid H = h)}{P(E = e)} \tag{1.23}$$

1.4 信息论的相关基本概念

1.4.1 Jensen 不等式

设 f 为区间上的凹函数，$p_i\in[0,1](i=1,2,\cdots,n)$，且 $\sum_{i=1}^{n} p_i=1$，则对任何 $x_i\in I$，均有

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \geqslant \sum_{i=1}^{n} p_i f(x_i) \tag{1.24}$$

若 f 为区间上严格凹函数，则式(1.24) 的等号只在下列条件满足时才成立：若 $p_i \cdot p_j\neq 0$，则必有 $x_i=x_j$。

1.4.2 熵

一个离散随机变量 X 的熵 $H(X)$的定义为

$$H(X)=\sum_{X}P(X)\mathrm{lb}\frac{1}{P(X)}=-\sum_{X}P(X)\mathrm{lb}P(X) \tag{1.25}$$

其中约定 $0\mathrm{lb}\frac{1}{0}=0$。通常选用 e 为对数的底。

熵是对随机变量的不确定性的度量，随机变量 X 的熵越大，说明它的不确定性也就越大。

1.4.3 联合熵、条件熵和互信息

联合熵是借助联合概率分布对熵的自然推广，两个离散随机变量 X 和 Y 的联合熵的定义为

$$H(X,Y)=\sum_{X,Y}P(X,Y)\mathrm{lb}\frac{1}{P(X,Y)}=-\sum_{X,Y}P(X,Y)\mathrm{lb}P(X,Y) \tag{1.26}$$

条件熵是利用条件概率分布对熵的一个延伸。随机变量 X 的熵是用它的概率分布 $P(X)$来定义的。如果知道另一个随机变量 Y 的取值为 y，那么 X 的后验分布即为 $P(X|Y=y)$。利用此条件分布可以定义给定 $Y=y$ 时 X 的条件熵为

$$H(X\mid Y=y)=\sum_{X}P(X\mid Y=y)\mathrm{lb}\frac{1}{P(X\mid Y=y)} \tag{1.27}$$

熵 $H(X)$度量的是随机变量 X 的不确定性，条件熵 $H(X|Y=y)$度量的则是已知 $Y=y$ 后，X 的不确定性。

对于式(1.28)，即

$$H(X,Y)=H(X)+H(Y\mid X)=H(Y)+H(X\mid Y) \tag{1.28}$$

称为熵的链规则。

在观测到 Y 以前，X 的不确定性是 $H(X)$，通过观测 Y，期望 X 的不确定性会变为 $H(X|Y)$，因此，$H(X)$与 $H(X|Y)$之差

$$I(X;Y)=H(X)-H(X\mid Y) \tag{1.29}$$

就是对 Y 包含多少 X 的信息的一个度量，称之为 Y 关于 X 的信息。可以证明，$I(X;Y)=I(Y;X)$成立，因此它又称为 X 和 Y 之间的互信息。

对任意两个离散随机变量 X 和 Y，有式(1.30)成立，即

$$I(X;Y)=\sum_{X,Y}P(X,Y)\mathrm{lb}\frac{P(X,Y)}{P(X)P(Y)} \tag{1.30}$$

对于 3 个离散随机变量，式(1.31)，即

$$I(X;Y\mid Z)=H(X\mid Z)-H(X\mid Z,Y) \tag{1.31}$$

称为给定 Z 时，Y 关于 X 的信息。容易证明 $I(X;Y|Z)=I(Y;X|Z)$，于是 $I(X;Y|Z)$也称为给定 Z 时，X 和 Y 之间的条件互信息。

1.4.4 相对熵

对定义于随机变量 X 的状态空间 Ω_x 上的两个概率分布 $P_1(X)$和 $P_2(X)$，可以用相对熵来度量它们之间的差异，即有

$$KL(P_1,P_2)=\sum_{X}P_1(X)\mathrm{lb}\frac{P_1(X)}{P_2(X)} \tag{1.32}$$

其中，$0\text{lb}\,\frac{0}{p_2}=0$，$p_1\text{lb}\,\frac{p_1}{0}=\infty$，$\forall p>0$。$KL(P_1,P_2)$又被称为$P_1(X)$和$P_2(X)$之间的Kullback-Leibler距离。但值得注意的是，$KL(P_1,P_2)\neq KL(P_2,P_1)$。

1.5 生成模型与判别模型

假设有观察值序列$Y=(Y_1,Y_2,\cdots,Y_n)$，求其对应的状态序列$X=(X_1,X_2,\cdots,X_n)$，则实际上就是求出状态序列X^*，使得条件概率$p(X_1,X_2,\cdots,X_n|Y_1,Y_2,\cdots,Y_n)$最大化，即

$$X^* = \arg\max_{X_1,X_2,\cdots,X_n} p(X_1,X_2,\cdots,X_n \mid Y_1,Y_2,\cdots,Y_n) \tag{1.33}$$

1. 生成模型

生成模型(Generative Models)不直接对$p(X_1,X_2,\cdots,X_n|Y_1,Y_2,\cdots,Y_n)$进行建模，而是先对其进行变换，构建联合概率$p(X_1,X_2,\cdots,X_n,Y_1,Y_2,\cdots,Y_n)$，即

$$\begin{aligned}p(X_1,X_2,\cdots,X_n \mid Y_1,Y_2,\cdots,Y_n) &= \frac{p(X_1,X_2,\cdots,X_n,Y_1,Y_2,\cdots,Y_n)}{p(Y_1,Y_2,\cdots,Y_n)}\\ &= \frac{p(Y_1,Y_2,\cdots,Y_n \mid X_1,X_2,\cdots,X_n) * p(X_1,X_2,\cdots,X_n)}{p(Y_1,Y_2,\cdots,Y_n)}\end{aligned} \tag{1.34}$$

在给定观察值序列的前提下，其出现的概率是一定的，所以式(1.33)可以由式(1.35)得出，即

$$\begin{aligned}X^* &= \arg\max_{X_1,X_2,\cdots,X_n} p(X_1,X_2,\cdots,X_n \mid Y_1,Y_2,\cdots,Y_n)\\ &= \arg\max_{X_1,X_2,\cdots,X_n} p(Y_1,Y_2,\cdots,Y_n \mid X_1,X_2,\cdots,X_n) * p(X_1,X_2,\cdots,X_n)\end{aligned} \tag{1.35}$$

生成模型认为观察值是由状态生成的。由生成模型的定义可知，隐马尔可夫模型是一种典型的生成模型。

2. 判别模型

判别模型(Discriminative Models)克服了生成模型的独立性假设，其直接对条件概率$p(X_1,X_2,\cdots,X_n|Y_1,Y_2,\cdots,Y_n)$进行建模，这就是说，在给定观察序列的条件下，寻找最可能的状态序列的时候，条件分布可以直接使用。

由于判别模型自身的特点，因此可以融合各种特征到模型中，使它成为一类适合序列标注的模型。常见的判别模型有最大熵模型MEM、条件随机场模型CRF等。显然，生成模型和判别模型有着明显的区别，下面对生成模型和判别模型进行对比。

(1) 统计建模方式不同。生成模型构建联合分布$p(X,Y)$；判别模型构建条件分布$p(X|Y)$。

(2) 训练时二者优化准则不同。生成模型优化训练数据的联合分布概率；判别模型优化训练数据的条件分布概率。判别模型与序列标记问题有较好的对应性。

(3) 对于观察序列的处理不同。生成模型中，观察序列作为模型的一部分；判别模型中，观察序列只作为条件，因此可以针对观察序列设计灵活的特征。

(4) 训练复杂度不同。判别模型训练复杂度较高。

(5) 是否支持无指导训练。生成模型支持无指导训练。

(6) 生成模型假设观察值之间具有严格的独立性,认为观察值是由状态生成的。不能融合各种特征和表达长距离依赖关系。而判别模型认为观察值(特征)决定状态,克服了生成模型的严格独立性假设,可以融合各种特征到模型中。

第2章 概率图模型的基本原理

2.1 概述

不同的概率图模型可以分成3类：有向概率图模型、无向概率图模型和混合概率图模型。有向概率图模型的典型代表是隐马尔可夫模型、贝叶斯网络和动态贝叶斯网络；无向概率图模型的典型代表是马尔可夫随机场和条件随机场；混合概率图模型的典型代表是链图。这些概率图模型刻画了随机变量间不同的条件独立关系。有向概率图模型通常用来表示随机变量间的因果关系，而无向概率图模型用来建立随机变量间的空间相互关系或者是相互依赖性。有向概率图模型和无向概率图模型只能表示随机变量的同一类关系，而混合概率图模型却可以表示不同类型的关系。

由于有向概率图模型和无向概率图模型被广泛使用，本章重点对这两种模型的基本原理、学习与推理算法进行阐述。

2.2 有向概率图模型

有向概率图模型(Directed Probabilistic Graphical Model)使用有向边连接不同的结点，这些有向边通常表示了结点间的因果关系。在有向概率图模型中，隐马尔可夫模型、贝叶斯网络和动态贝叶斯网络被广泛的使用。

2.2.1 隐马尔可夫模型

隐马尔可夫模型(Hidden Markov Model,HMM)是20世纪60年代末70年代初提出来的一种基于马尔可夫源或可尔可夫链概率函数的统计信号模型，它是一种用参数表示的，用于描述随机过程统计特性的概率模型。隐马尔可夫模型经常被用作建立随时间变化的随机变量概率分布模型。

HMM的理论基础是由Baum等人建立起来的，随后由CMU的Baker和IBM的Jelinet等将其应用到语音识别领域。由于Bell实验室的Rabiner等在20世纪80年代中期对HMM的深入浅出的介绍，才逐渐使HMM被世界各国研究人员所了解和熟悉，进而成为公认的一个研究热点。特别是在近些年HMM可以为不同信号建模的特点，以及在语音信号处理上的成功，并且拥有经典的训练和寻优算法，使得很多研究人员开始尝试将HMM用于目标识别、人脸识别及OCR等领域。

1. HMM的理论基础——马尔可夫链

Markov(马尔可夫,1856—1922),俄国数学家,师从大数学家切比雪夫,执教于圣彼得堡大学,1896年被选为圣彼得堡科学院院士。1907年马尔可夫提出了“马尔可夫链”——一种能用数学分析方法研究自然过程的一般方式,开创了对一种无后效性的随机过程——“马尔可夫过程”的研究。马尔可夫链是马尔可夫随机过程的特殊情况,即马尔可夫链是状态和时间参数都离散的马尔可夫过程。从数学上,可以给出以下定义:

随机序列 X_n 在任一时刻 n,它可以处在状态 $\theta_1,\cdots,\theta_N$,且它在 $m+k$ 时刻所处的状态为 q_{m+k} 的概率,只与它在 m 时刻的状态 q_m 有关,而与 m 时刻以前它所处状态无关,即有

$$\begin{aligned}P(X_{m+k} = q_{m+k} \mid X_m = q_m, X_{m-1} = q_{m-1}, \cdots, X_1 = q_1) \\ = P(X_{m+k} = q_{m+k} \mid X_m = q_m)\end{aligned} \tag{2.1}$$

其中,$q_1,q_2,\cdots,q_m,q_{m+k}\in(\theta_1,\theta_2,\cdots,\theta_N)$,则称 X_m 为马尔可夫链,并且称

$$P_{ij}(m,m+k) = P(q_{m+k} = \theta_j \mid q_m = \theta_i) \tag{2.2}$$

为 k 步转移概率。其中,$1\leqslant i,j\leqslant N$,$m$、$k$ 为正整数。当 $P_{ij}(m,m+k)$ 与 m 无关时,称这个马尔可夫链为齐次马尔可夫链,此时有

$$P_{ij}(m,m+k) = P_{ij}(k) \tag{2.3}$$

以后若无特别声明,马尔可夫链就是指齐次马尔可夫链。当 $k=1$ 时,$P_{ij}(1)$ 称为一步转移概率,简称为转移概率,记为 a_{ij}。所有转移概率 $a_{ij}(1\leqslant i,j\leqslant N)$ 可以构成一个转移概率矩阵,即

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \tag{2.4}$$

且有

$$0 \leqslant a_{ij} \leqslant 1, \quad \sum_{j=1}^{N} a_{ij} = 1$$

由于 k 步转移概率 $P_{ij}(k)$ 可由转移概率 a_{ij} 得到,因此,描述马尔可夫链的最重要参数就是转移概率矩阵 $\boldsymbol{A}$。但 $\boldsymbol{A}$ 矩阵还决定不了初始分布,即由 $\boldsymbol{A}$ 求不出 $q_1=\theta_i$ 的概率,这样,完全描述马尔可夫链,除 $\boldsymbol{A}$ 矩阵之外,还必须引进初始概率矢量 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)$,其中

$$\pi_i = P(q_1 = \theta_i) \quad 1 \leqslant i \leqslant N \tag{2.5}$$

显然有

$$0 \leqslant \pi_i \leqslant 1, \quad \sum_i \pi_i = 1$$

实际中,马尔可夫链的每一状态可以对应于一个可观测到的物理事件。比如天气预测中的雨、晴、雪等,那么,这时它可称为天气预报的马尔可夫链模型。根据这个模型,可以算出各种天气(状态)在某一时刻出现的概率。

2. HMM概念

HMM是在马尔可夫链的基础上发展起来的。由于实际问题比马尔可夫链模型所描述的更为复杂,观察到的事件并不是与状态一一对应,而是通过一组概率分布相联系,这样的

模型就称为 HMM。它是一个双重随机过程，其中之一是马尔可夫链，这是基本随机过程，它描述状态的转移。另一个随机过程描述状态和观察值，不像马尔可夫链模型中的观察值和状态一一对应，因此，不能直接看到状态，而是通过一个随机过程去感知状态的存在及其特性。因而称之为“隐”马尔可夫模型，即 HMM。

例如，著名的球和缸(Ball and Urn)实验，就很好地说明了隐马尔可夫模型。

设有 N 个缸，每个缸中装有很多颜色的球，球的颜色由一组概率分布描述。实验是这样进行的，根据某个初始概率分布，随机地选择 N 个缸中的一个，如第 i 个缸，再根据这个缸中彩色球颜色的概率分布，随机地选择一个球，记下球的颜色，记为 O_1，再把球放回缸中，又根据描述缸的转移的概率分布，随机选择下一个缸。例如，第 j 个缸，再从缸中随机选一个球，记下球的颜色，记为 O_2，一直进行下去。可以得到一个描述球的颜色的序列 O_1，O_2，…，由于这是观察到的事件，因而称之为观察值序列。但缸之间的转移以及每次选取球的缸被隐藏起来了，并不能直接观察到。而且，从每个缸中选取球的颜色并不是与缸一一对应，而是由该缸中彩球颜色概率分布随机决定的。此外，每次选取哪个缸则由一组转移概率所决定。球与缸试验如图 2.1 所示。

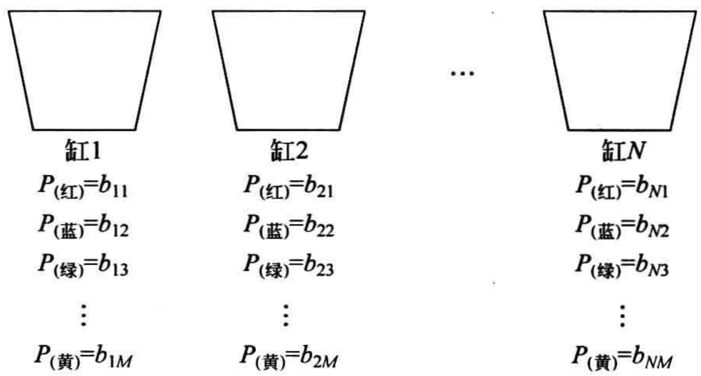


图 2.1 球和缸试验

下面给出 HMM 的定义，或者说，一个 HMM 可以由下列参数描述：

(1) 模型中马尔可夫链状态数目 N：记 N 个状态为 $\theta_1,\cdots,\theta_N$，记 t 时刻马尔可夫链所处状态为 q_t，显然 q_t 属于$(\theta_1,\theta_2,\cdots,\theta_N)$。在球与缸实验中的缸就相当于状态。

(2) 每个状态对应的可能的观察值数目 M：记 M 个观察值为 $V_1,\cdots,V_M$，记 t 时刻观察到的观察值为 O_t，其中 O_t 属于$(V_1,\cdots,V_M)$。在球与缸实验中所选彩球的颜色，就是观察值。

(3) 初始状态概率矢量$\boldsymbol{\pi}$：$\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)$，其中

$$\pi_i = P(q_1 = \theta_i) \quad 1 \leqslant i \leqslant N \tag{2.6}$$

在球与缸实验中指开始时选取某个缸的概率。

(4) 状态转移概率矩阵 $\boldsymbol{A}$：$\boldsymbol{A}=(a_{ij})_{N\times N}$，其中

$$a_{ij} = P(q_{t+1} = \theta_j \mid q_t = \theta_i) \quad 1 \leqslant i,j \leqslant N \tag{2.7}$$

在球与缸实验中指描述每次在当前选取的缸的条件下选取下一个缸的概率。

(5) 观察值概率矩阵 $\boldsymbol{B}$：假设观测变量的样本空间为 V，在状态 θ_j 时输出观测变量的概率分布可表示为 $\boldsymbol{B}=(b_{jk})_{N\times M}$，其中

$$b_{jk} = P(O_t = V_k \mid q_t = \theta_j) \quad 1 \leqslant j \leqslant N;\ 1 \leqslant k \leqslant M \tag{2.8}$$

其中，O_t 为时刻 t 的观测随机变量，可以是一个数值或向量，观测序列记为 $O=\{O_1,O_2,\cdots,O_t\}$。值得注意的是，此处观测变量的样本空间和概率分布可以为离散型，也可以为连续型。在球与缸实验中，b_{jk} 就是第 j 个缸中球的颜色 k 出现的概率。

综上可知，要描述一个完整的 HMM，需要模型参数($N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B}$)。为了简化，常用下面的形式来表示，即

$$\lambda = (N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B}) \tag{2.9}$$

或简写为

$$\lambda = (\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B}) \tag{2.10}$$

图 2.2 可以更形象地说明隐马尔可夫模型的定义。HMM 可分为两部分，一个是马尔可夫链，由$\boldsymbol{\pi}$、$\boldsymbol{A}$ 描述，产生的输出为状态序列，另一个是一个随机过程，由 $\boldsymbol{B}$ 描述，产生的输出为观察值序列，如图 2.2 所示。T 为观察值时间长度。

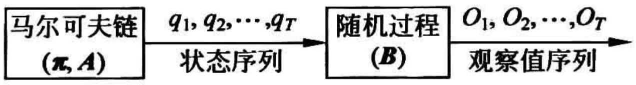


图 2.2 HMM组成示意图

图 2.3 出示了一个简单的第一定律隐马尔可夫模型，图中 X 是隐状态变量，Y 是观测变量，观测结点提供了相应隐结点的基本的局部信息。隐马尔可夫模型假设隐状态随机变量的分布形成了马尔可夫链。对于第一定律隐马尔可夫模型，当给定父结点，隐状态变量是条件独立于其他父结点。基于模型中给定的条件独立关系，所有结点的联合概率公式可表示为

$$P(\{X_i,Y_i\}_{i=1}^{T}) = P(X_1)P(Y_1 \mid X_1)\prod_{i=2}^{T}[P(X_i \mid X_{i-1})P(Y_i \mid X_i)] \tag{2.11}$$

式中，$P(X_i|X_{i-1})$为各时间片隐结点间的转移概率；$P(Y_i|X_i)$为观测结点的似然概率；$P(X_1)$为 X_1 结点的先验概率。

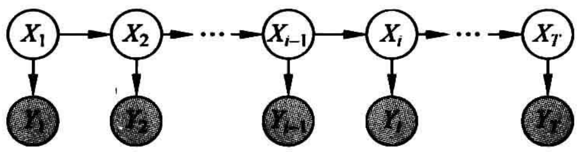


图 2.3 一个简单的第一定律隐马尔可夫模型

3. HMM 基本算法

标准隐马尔可夫模型在实际应用中要解决 3 个基本问题，即评估问题、解码问题和学习问题。

问题 1 给定观察序列 $O=O_1,O_2,\cdots,O_T$和$\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$，计算 $P(O/\lambda)$，即给定模型和输出观察序列，如何计算从模型生成观察序列的概率。可以把它看作是评估一个模型和给定观察输出序列的匹配程度，由此可以用来在一系列候选对象中选取最佳的匹配。

问题 2 给定观察序列 $O=O_1,O_2,\cdots,O_T$ 和模型λ，求在某种有意义的情况下最优的相关状态序列 $Q^*=q_1^*,q_2^*,\cdots,q_T^*$，即最好地解释观察情况。该问题可以理解为对输出观察的最佳“解释”，它试图揭示模型的隐藏部分，比如说查找“正确”的状态序列，在应用中，通常

都使用一个优化策略来最大可能地解决这个问题。

问题3 怎样调整模型参数$\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$，使$P(O/\lambda)$最大？它试图优化模型的参数来最佳地描述一个给定的观察序列是如何得来的。

这三大基本问题的解决过程与方法，即是HMM实际应用的过程与方法。下面针对以上3个问题，给出HMM的3个基本算法描述。

1）前向-后向算法

这个算法是用来计算给定一个观察值序列$O=O_1,O_2,\cdots,O_T$及一个模型$\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$时，由模型$\lambda$产生出$O$的概率$P(O/\lambda)$。

$P(O/\lambda)$最直接的求取方法如下：

对一个固定的状态序列$S=q_1,q_2,\cdots,q_T$，有

$$P(O/S,\lambda)=\prod_{t=1}^{T}P(O_t/q_t,\lambda)=b_{q_1}(O_1)b_{q_2}(O_2)\cdots b_{q_T}(O_T) \tag{2.12}$$

其中

$$b_{q_t}(O_t)=b_{jk}\mid q_t=\theta_j,O_t=V_k\quad 1\leqslant t\leqslant T \tag{2.13}$$

而对给定的λ，产生S的概率为

$$P(S/\lambda)=\pi_{q_1}a_{q_1q_2}\cdots a_{q_{T-1}q_T} \tag{2.14}$$

因此，所求概率为

$$P(O/\lambda)=\sum_{S}P(O/S,\lambda)P(S/\lambda)$$
$$=\sum_{q_1,q_2,\cdots,q_T}\pi_{q_1}b_{q_1}(O_1)a_{q_1q_2}b_{q_2}(O_2)\cdots a_{q_{T-1}q_T}b_{q_T}(O_T) \tag{2.15}$$

显而易见，式(2.15)的计算量是十分惊人的，大约为$2TN^T$数量级，当$N=5,T=100$时，计算量达10^{72}，这是完全不能接受的。在此情况下，要求出$P(O/\lambda)$还必须寻求更有效的算法，这就是Baum等提出的前向-后向算法。

(1) 前向算法。定义前向变量为

$$\boldsymbol{\alpha}(i)=P(O_1,O_2,\cdots,O_t,q_t=\theta_i\mid\lambda)\quad 1\leqslant t\leqslant T \tag{2.16}$$

那么，有以下过程。

a. 初始化，即

$$\alpha_1(i)=\pi_i b_i(O_1)\quad 1\leqslant i\leqslant N \tag{2.17}$$

b. 递归，即

$$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(O_{t+1})\quad 1\leqslant t\leqslant T-1;\ 1\leqslant j\leqslant N \tag{2.18}$$

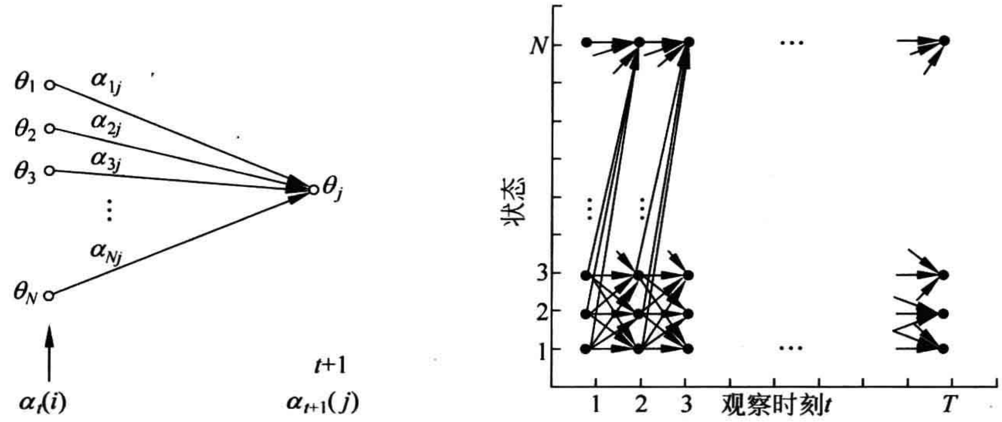
c. 终结，即

$$P(O/\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{2.19}$$

其中

$$b_j(O_{t+1})=b_{jk}\mid_{O_{t+1}=V_k} \tag{2.20}$$

这种算法计算量大为减少，变为$N(N+1)(T-1)+N$次乘法和$N(N-1)(T-1)$次加法。同样，$N=5,T=100$时，只需大约3000次计算(乘法)。这种算法是一种典型的栅格结构，如图2.4所示。



(a) t时刻递归关系　　(b) 栅格结构

图 2.4　前向算法示意图

(2) 后向算法。与前向算法类似，定义后向变量为

$$\boldsymbol{\beta}_t(i) = P(O_{t-1}, O_{t-2}, \cdots, O_T \mid q_t = \theta_i, \lambda) \quad 1 \leqslant t \leqslant T-1 \tag{2.21}$$

其中，$\beta_T(i)=1$。

类似地，有以下过程。

a. 初始化，即

$$\beta_T(i) = 1 \quad 1 \leqslant i \leqslant N \tag{2.22}$$

b. 递归，即

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, \cdots, 1;\ 1 \leqslant i \leqslant N \tag{2.23}$$

c. 终结，即

$$P(O \mid \lambda) = \sum_{i=1}^{N} \beta_1(i) \tag{2.24}$$

后向算法的计算量大约在 N^2T 数量级，也是一种栅格结构。

2) Viterbi 算法

这个算法解决了给定一个观察值序列 $O=O_1, O_2, \cdots, O_T$ 和一个模型 $\lambda=(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$，在最佳的意义上确定一个状态序列 $\boldsymbol{Q}^*=q_1^*, q_2^*, \cdots, q_T^*$ 的问题。

“最佳”的意义有很多种，由不同的定义可得到不同的结论。这里讨论的最佳意义上的状态序列 $\boldsymbol{Q}^*$，是指使 $P(Q,O|\lambda)$ 最大时确定的状态序列 $\boldsymbol{Q}^*$。Viterbi 算法可以叙述如下：

定义 $\delta_t(i)$ 为时刻 t 时沿一条路径 $q_1, q_2, \cdots, q_t$，且 $q_t=\theta_i$，产生出 $O_1, O_2, \cdots, O_t$ 的最大概率，即有

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P(q_1, q_2, \cdots, q_t, q_t = \theta_i, O_1, O_2, \cdots, O_t \mid \lambda) \tag{2.25}$$

那么，求取最佳状态序列 $\boldsymbol{Q}^*$ 的过程如下。

a. 初始化，即

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \leqslant i \leqslant N \tag{2.26}$$

$$\varphi_1(i) = 0 \quad 1 \leqslant i \leqslant N \tag{2.27}$$

b. 递归，即

$$\delta_t(j) = \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) a_{ij}] b_j(O_t) \quad 2 \leqslant t \leqslant T;\ 1 \leqslant j \leqslant N \tag{2.28}$$

$$\varphi_t(j) = \arg\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}] \quad 2\leqslant t\leqslant T;\ 1\leqslant j\leqslant N \tag{2.29}$$

c. 终结，即

$$P^* = \max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{2.30}$$

$$q_T^* = \arg\max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{2.31}$$

d. 状态序列求取，即

$$q_t^* = \varphi_{t+1}(q_{t+1}^*) \quad t = T-1, T-2, \cdots, 1 \tag{2.32}$$

应当指出，Viterbi 算法的一个副产品 $P^* = \max\limits_Q P(Q,O|\lambda)$ 和前向-后向算法计算出的 $P(O|\lambda) = \sum\limits_Q P(Q,O \mid \lambda)$ 之间的关系为：$P(Q,O|\lambda)$ 动态范围很大，或者说不同的 Q 使 $P(Q,O|\lambda)$ 的值差别很大，而 $\max\limits_Q P(Q,O|\lambda)$ 事实上是 $\sum\limits_Q P(Q,O|\lambda)$ 中举足轻重的唯一成分，因此，常常等价地使用 $\max\limits_Q P(Q,O|\lambda)$ 和 $\sum\limits_Q P(Q,O \mid \lambda)$，那么，Viterbi 算法也能用来计算 $P(O|\lambda)$。

此外，上述的 Viterbi 算法也是一种栅格结构，而且类似于前向算法。同样，由后向算法的思想出发，亦可推导出 Viterbi 算法的另一种实现方式。

3) Baum-Welch 算法

这个算法实际上是解决 HMM 训练，即 HMM 参数估计问题，或者说给定一个观察值序列 $O=O_1, O_2, \cdots, O_T$，该算法能确定一个 $\lambda=(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$，使 $P(O/\lambda)$ 最大。

显然，由式(2.16)和式(2.21)定义的前向和后向变量，有

$$P(O \mid \lambda) = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j) \quad 1\leqslant t\leqslant T-1 \tag{2.33}$$

这里，求取 λ，使 $P(O|\lambda)$ 最大，是一个泛函极值问题。但是，由于给定的训练序列有限，因而不存在一个最佳的方法来估计 λ。在这种情况下，Baum-Welch 算法利用递归的思想，使 $P(O|\lambda)$ 局部极大，最后得到模型参数 $\lambda=(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$。此外，用梯度方法也可以达到类似目的。

定义 $\xi_t(i,j)$ 为给定训练序列 O 和模型 λ 时，时刻 t 时马尔可夫链处于 θ_i 状态和时刻 $t+1$ 为 θ_j 状态的概率，即

$$\xi_t(i,j) = P(O, q_t = \theta_i, q_{t+1} = \theta_j \mid \lambda) \tag{2.34}$$

可以推导出

$$\xi_t(i,j) = [\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)]/P(O \mid \lambda) \tag{2.35}$$

那么，时刻 t 时马尔可夫链处于 θ_i 状态的概率为

$$\xi_t(i) = P(O, q_t = \theta_i \mid \lambda) = \sum_{j=1}^{N}\xi_t(i,j) = \alpha_t(i)\beta_t(i)/P(O \mid \lambda) \tag{2.36}$$

因此，$\sum\limits_{t=1}^{T-1}\xi_t(i)$ 表示从 θ_i 状态转移出去的次数的期望值，而 $\sum\limits_{t=1}^{T-1}\xi_t(i,j)$ 表示从 θ_i 状态转移到 θ_j 状态的次数的期望值。由此，导出了 Baum-Welch 算法中著名的重估 (Reestimation) 公式，即

$$\bar{\pi}_i = \xi_1(i) \tag{2.37}$$

$$\bar{a}_{ij} = \sum_{t=1}^{T-1}\xi_t(i,j) \Big/ \sum_{t=1}^{T-1}\xi_t(i) \tag{2.38}$$

$$\bar{b}_{jk} = \sum_{t=1 且 O_t=V_k}^{T} \xi_t(j) \Big/ \sum_{t=1}^{T} \xi_t(j) \tag{2.39}$$

那么，HMM 参数 $\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$ 的求取过程为：根据观察值序列 O 和选取的初始模型 $\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$，由重估式(2.37)、式(2.38)和式(2.39)，求得一组新参数 $\bar{\pi}_i$、$\bar{a}_{ij}$ 和 $\bar{b}_{jk}$，亦即得到了一个新的模型 $\bar{\lambda}=(\bar{\boldsymbol{\pi}},\bar{\boldsymbol{A}},\bar{\boldsymbol{B}})$，可以证明，$P(O|\bar{\lambda})>P(O|\lambda)$，即由重估公式得到的 $\bar{\lambda}$ 比 λ 在表示观察值序列 O 方面要好。那么，重复这个过程，逐步改进模型参数，直到 $P(O|\bar{\lambda})$ 收敛，即不再明显增大，此时的 $\bar{\lambda}$ 即为所求的模型。

应当指出，HMM 训练或称参数估计问题，与前面讨论的两个问题相比，是最困难的一个问题，Baum-Welch 算法只是得到广泛应用的解决这一问题的经典方法，但并不是唯一的，也远不是最完善的方法。

4. HMM 算法实现中的基本问题

1）初始模型选取

根据 Baum-Welch 算法由训练数据得到 HMM 参数时，如重估式(2.37)、式(2.38)和式(2.39)，一个重要问题就是初始模型的选取。不同的初始模型将产生不同的训练结果。因为算法是使 $P(O|\lambda)$ 局部极大时得到的模型参数，因此，选取好的初始模型，使最后求出的局部极大与全局最大接近，是很有意义的。

但是，至今这个问题仍没有完美的答案。实际处理时都是采用一些经验方法。一般认为，$\boldsymbol{\pi}$ 和 $\boldsymbol{A}$ 参数初值选取影响不大，可以随机选取或均匀取值，只要满足 $0\leqslant a_{ij}\leqslant 1$、$\sum_{j=1}^{N} a_{ij}=1$、$0\leqslant \pi_i\leqslant 1$、$\sum_{i} \pi_i=1$ 要求的约束条件即可。但 $\boldsymbol{B}$ 的初值对训练出的 HMM 影响较大，一般倾向采取较为复杂的初值选取方法。基于这种考虑，一种典型的 HMM 参数估计过程如图 2.5 所示。

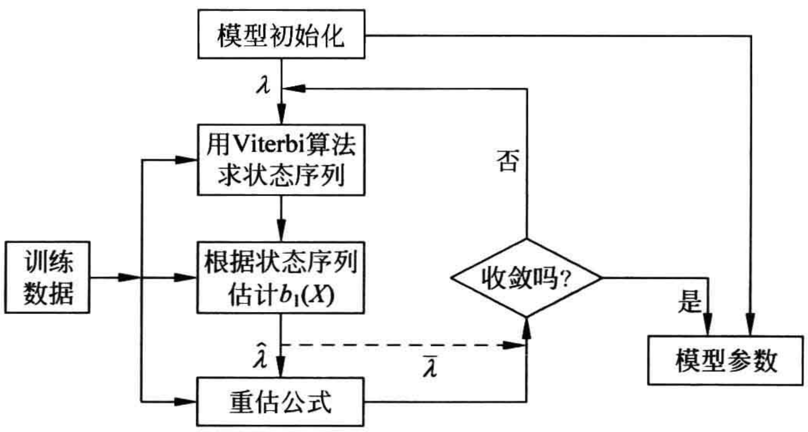


图 2.5 一种 HMM 参数估计方法示意图

这里，初始模型 λ 可以任意选取。但因为有 $P(O|\hat{\lambda})>P(O|\lambda)$，所以 $\hat{\lambda}$ 是 λ 改进后的模型。再将 $\hat{\lambda}$ 作为初始值用重估公式，得到 $\bar{\lambda}$。这样就避免了初值的选择不当，变经典的 $\lambda\rightarrow\bar{\lambda}$ 为 $\lambda\rightarrow\hat{\lambda}\rightarrow\bar{\lambda}$。当然，沿图 2.5 中虚线，不用重估公式，$\hat{\lambda}$ 也可近似作为模型参数。

2）多个观察值序列训练

实际中，训练一个HMM，经常是用到不止一个观察值序列，那么，对于L个观察值序列训练HMM时，要对Baum-Welch算法的重估公式(2.37)、式(2.38)和式(2.39)加以修正。设L个观察秩序列为$O^{(l)}(l=1,\cdots,L)$，其中$O^{(l)}=O_1^{(l)},O_2^{(l)},\cdots,O_{T_1}^{(l)}$，假定各个观察值序列独立，此时有

$$P(O\mid\lambda)=\prod_{l=1}^{L}P(O^{(l)}\mid\lambda) \tag{2.40}$$

由于重估公式是以不同事件的频率为基础的，因此，对L个训练序列，重估公式修正为

$$\bar{\pi}_i=\sum_{l=1}^{L}\alpha_1^{(l)}(i)\beta_1^{(l)}(i)/P(O^{(l)}\mid\lambda)\quad 1\leqslant i\leqslant N \tag{2.41}$$

$$\bar{a}_{ij}=\frac{\sum_{l=1}^{L}\sum_{t=1}^{T-1}\alpha_t^{(l)}(i)a_{ij}b_j(O_{t+1}^{(l)})\beta_{t+1}^{(l)}(j)/P(O^{(l)}\mid\lambda)}{\sum_{l=1}^{L}\sum_{t=1}^{T_t}\alpha_t^{(l)}(i)\beta_t^{(l)}(i)/P(O^{(l)}\mid\lambda)}\quad 1\leqslant i;\ j\leqslant N \tag{2.42}$$

$$\bar{b}_{jk}=\frac{\sum_{l=1}^{L}\sum_{\substack{t=1\\ \text{且}O_t=V_k}}^{T_t}\alpha_t^{(l)}(j)\beta_t^{(l)}(j)/P(O^{(l)}\mid\lambda)}{\sum_{l=1}^{L}\sum_{t=1}^{T_t}\alpha_t^{(l)}(i)\beta_t^{(l)}(i)/P(O^{(l)}\mid\lambda)}\quad 1\leqslant j\leqslant N;\ 1\leqslant k\leqslant M \tag{2.43}$$

3）比例因子问题

在前向-后向算法和Baum-Welch算法中，都有$\alpha_t(i)$和$\beta_t(i)$的递归计算，因为所有量都小于1，因此，$\alpha_t(i)$（随着t的增加）和$\beta_t(i)$（随着t的减少）都迅速趋向于零，为了解决这种下溢(Underflow)问题，必须采取增加比例因子(Scaling)的方法，对有关算法加以修正，处理过程如下。

(1) 对α的处理。

$$\alpha_1(i)=\pi_i b_i(O_1)\quad 1\leqslant i\leqslant N \tag{2.44}$$

$$\alpha_1^*(i)=\frac{\alpha_1(i)}{\sum_{i=1}^{N}\alpha_1(i)}\overset{\Delta}{=}\frac{\alpha_1(i)}{\Phi_1}\quad 1\leqslant i\leqslant N \tag{2.45}$$

$$\bar{\alpha}_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t^*(i)a_{ij}\right]b_j(O_{t+1})\quad 1\leqslant j\leqslant N;\ t=1,2,\cdots,T-1 \tag{2.46}$$

$$\alpha_{t+1}^*(j)=\bar{\alpha}_{t+1}(j)/\sum_{j=1}^{N}\bar{\alpha}_{t+1}(j)\overset{\Delta}{=}\bar{\alpha}_{t+1}(j)/\Phi_{t+1}\quad 1\leqslant j\leqslant N;\ t=1,2,\cdots,T-1 \tag{2.47}$$

(2) 对β的处理。

$$\beta_T(i)=1\quad 1\leqslant i\leqslant N \tag{2.48}$$

$$\beta_T^*(i)=1\quad 1\leqslant i\leqslant N \tag{2.49}$$

$$\tilde{\beta}_t(i)=\sum_{j=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}^*(j)\quad 1\leqslant i\leqslant N;\ t=T-1,\cdots,1 \tag{2.50}$$

$$\beta_t^*(i)=\tilde{\beta}_t(i)/\Phi_{t+1}\quad 1\leqslant i\leqslant N;\ t=T-1,\cdots,1 \tag{2.51}$$

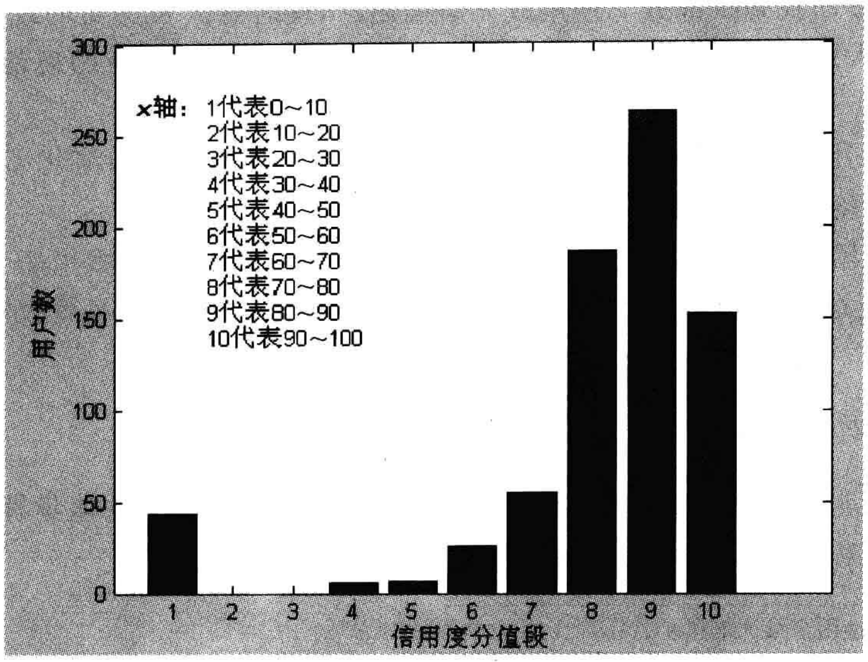


图 8.6　客户信用度分布

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	3	1	2	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	3	1	1	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	3	1	2	2	1	1	2	2	3	1	1	5	2	2	2	50.378
1	1	3	1	2	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	1	2	1	1	1	1	3	1	1	5	2	2	2	82.364
1	1	3	1	2	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	1	2	1	1	1	1	1	1	1	4	2	2	2	90.606
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	5	1	2	2	1	1	1	1	3	1	1	5	2	2	2	82.364
2	2	2	1	2	2	1	2	1	1	3	1	1	4	2	1	1	36.966
1	1	3	2	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	1	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	1	2	1	1	1	1	1	1	3	2	2	2	2	100
1	1	2	1	2	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	2	1	1	2	1	1	1	1	1	1	2	4	2	2	2	79.945
1	1	3	1	1	2	1	1	1	1	3	1	2	3	2	2	2	73.266
1	1	2	1	1	2	1	1	1	1	1	1	1	5	2	1	2	100
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	2	2	1	2	1	1	3	1	1	4	2	2	2	77.08
1	2	3	1	1	2	1	1	1	1	1	1	1	3	2	2	2	91.348
1	1	2	1	2	2	1	1	1	1	3	2	1	3	2	2	2	100
1	1	2	1	1	2	1	1	1	1	1	1	1	4	2	1	2	90.606

图 8.7　信用度评分模型预测结果

8.2.3　基于投票熵和最小相对熵的主动贝叶斯网络建立客户信用等级评估模型

1. 问题背景

客户信用风险等级是电信企业决策支持系统中进行客户总量分析、客户新增与流失分析、客户行为分析、经营风险分析和客户欺诈分析的一个重要分析角度，其可以应用于各种信用风险管理和控制。准确地评估客户信用等级，可以解决在复杂的用户消费群体中良莠

连续 HMM 可以认为是 HMM 的一种一般形式,连续 HMM 和离散 HMM 都是其特例。而且,半连续 HMM 在一定程度上兼有二者的长处。因为,HMM 对连续矢量使用离散 HMM 信息丢失较大,但是如果使用连续的 HMM,需要使用较多的概率密度函数进行混合,模型复杂,运算量大,并且需要使用更多的训练数据才能得到可靠的模型。如果对图 2.2 所示的第一部分,也就是由$\boldsymbol{\pi}$、$\boldsymbol{A}$ 描述的马尔可夫链加以修正,就可以得到另外 3 种主要的 HMM,即:利用 Gibbs 分布取代马尔可夫链的 HMM;在马尔可夫链中加入状态驻留时间参数的 HMM;二阶 HMM。此外,还有其他一些形式的 HMM 存在于各个应用领域,在此不再赘述。

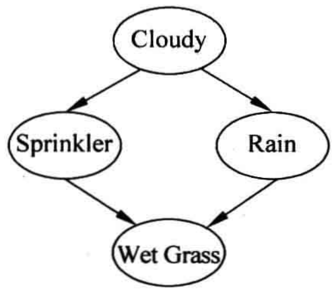
2.2.2 贝叶斯网络

1. 概念

贝叶斯网络(Bayesian Network,BN)一般是指带有概率信息的有向无环图(Directed Acyclic Graph,DAG)。贝叶斯网络的信息由两部分组成:首先是表示条件独立性信息的一种自然方式——网络结构 S,S 中的每一结点表示特定域中的一个概念或变量,在结点间的连接(有向弧)表示了可能的因果关系,体现了域知识定性方面的特征;其次,每一结点都附有与该变量相联系的条件概率分布函数(Conditional Probability Distribution, CPD),如果变量是离散的,则它表现为给定其父母结点状态时该结点取不同值的条件概率表(Conditional Probability Table, CPT)。CPT 体现了域知识定量方面的特征。可见,贝叶斯网络是一种表示数据变量间潜在关系的定性定量的方法,它使用这种图形结构指定了一组条件独立的声明和用于刻画概率依赖强度的条件概率的数字值。图 2.6 显示了一个简单的贝叶斯网络。

$P(C$=F)	$P(C$=T)
0.5	0.5

C	$P(S$=F)	$P(S$=T)
F	0.5	0.5
T	0.9	0.1



C	$P(R$=F)	$P(R$=T)
F	0.8	0.2
T	0.2	0.8

S	R	$P(W$=F)	$P(W$=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

图 2.6 一个贝叶斯网络的例子

由于贝叶斯网络表示了因果过程的总体结构,故它可被看作是拥有许多不同组合的一个抽象知识库。它的语义可以从两方面来理解:一方面是将网络看作一种联合概率分布的

表示,即 BNs 完整、紧凑地表示了网络中各变量的联合概率分布;另一方面将网络看作条件独立性声明集合的一种表示。这两种观点实质上是等价的,而且在贝叶斯网络的表达、学习、推理算法中都得到统一。假设一组有限集合$\{Y_1,\cdots,Y_n\}$表示一组离散随机变量,它们分别取值$\{y_1,\cdots,y_n\}$的联合概率为

$$\begin{aligned}P(y_1,\cdots,y_n) &= P(y_n \mid y_{n-1},\cdots,y_1)P(y_{n-1} \mid y_{n-2},\cdots,y_1)\cdots P(y_2 \mid y_1)P(y_1)\\ &= \prod_{i=1}^{n} P(y_i \mid y_{i-1},\cdots,y_1)\end{aligned} \tag{2.61}$$

在不确定信息领域,条件独立性是一种构造知识重要的、鲁棒的方法。在贝叶斯网络中,人们断言每一结点在给定其父母结点后都条件独立于它的前辈结点,故有

$$P(y_1,\cdots,y_n) = \prod_{i=1}^{n} P(y_i \mid \mathrm{Pa}(y_i)) \tag{2.62}$$

式中,$\mathrm{Pa}(y_i)\subseteq\{y_{i-1},\cdots,y_1\}$是结点 y_i 的父母结点组,它们的取值已知,可见在联合概率空间中的每一状态都可用贝叶斯网络中条件概率表的适当元素的乘积来表示。

当贝叶斯网络作为分类器时,设某领域中类别空间为 $C=\{c_1,c_2,\cdots,c_l\}$,特征空间为 $\boldsymbol{X}=<X_1,X_2,\cdots,X_m>$,每个特征变量的值域为 $\mathrm{Val}(X_i)(i\in[1\cdots m])$,特征变量的取值用小写字母 $x_i(i\in[1\cdots m])$表示。对某一实例 $\boldsymbol{x}=<x_1,x_2,\cdots,x_m>$来说,分类的目的就是通过学习一定的训练样本集 D,来获得它的类别标注 c。贝叶斯网络分类器采用表达式 $\max\limits_{i=[1\cdots l]}\{p(c_i|\boldsymbol{x})\}$来决定它的类别,其中:

$$p(c_i \mid \boldsymbol{x}) = \frac{p(c_i)\times\prod_{j=1}^{m} p(x_j \mid c_i;\pi(x_j))}{p(\boldsymbol{x})} \tag{2.63}$$

式中,$\pi(x_j)$为结点 X_j 除类别结点 C 之外的所有父结点;x_j 为实例 $\boldsymbol{x}$ 第 j 个特征的取值。所以,学习贝叶斯网络分类的任务是从训练样本集 D 中学习概率分布函数:$p(c_i)$,$p(x_j|c_i;\pi(x_j))(i\in[1\cdots l],j\in[1\cdots m])$。

在这个学习任务中,包括两个过程:一是对于每一特征结点找到除类别结点之外的所有父结点,也就是学习贝叶斯网络的结构;二是在已知结构的基础上,获得这些参数的估计,即参数学习问题。

2. 几种常用贝叶斯网络分类器

1) 朴素贝叶斯网络

朴素贝叶斯网络(Naive Bayesian Networks, NBN)是贝叶斯网络中最简单的一种。它将类别结点作为根结点,其各属性结点相互独立,且都以类别结点为父结点。其结构如图 2.7 所示。

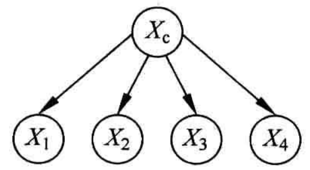


图 2.7 朴素贝叶斯网络

采用朴素贝叶斯分类器,分类任务即转化成了:已知 X_1,$X_2,\cdots,X_n$ 的值,求 $p(X_c|X_1,X_2,\cdots,X_n)$最大时 X_c 的取值问题,根据贝叶斯公式,即

$$p(X_{cj} \mid X_1X_2\cdots X_n) = \frac{p(X_1X_2\cdots X_n \mid X_{cj})p(X_{cj})}{p(X_1X_2\cdots X_n)} = \frac{p(X_1X_2\cdots X_n \mid X_{cj})p(X_{cj})}{\sum_{i=1}^{k} p(X_1X_2\cdots X_n \mid X_{ci})p(X_{ci})} \tag{2.64}$$

式中，$p(X_{cj})$为先验概率。

由于 $X_1, X_2, \cdots, X_n$ 相互独立，根据链规则有

$$p(X_1 X_2 \cdots X_n \mid X_{cj}) = \prod_{i=1}^{n} p(X_i \mid X_{cj}) \tag{2.65}$$

由以上公式求得后验概率，比较后验概率的大小，可求得最大后验概率时对应的 X_{cj} 值，即完成分类任务。

朴素贝叶斯分类器虽然简单，但在实验中表现出的分类能力与经典的 C4.5 分类器不相上下。然而，由于其对属性变量间的独立性要求较强，在实际中较难得到满足。

2）通用贝叶斯网络

通用贝叶斯网络(General Bayesian Networks，GBN)是将类结点和属性结点作为同等地位的网络结点，根据数据集中的数据训练出贝叶斯网络，直接作为分类器的。用通用贝叶斯网络分类器进行分类的过程，实际上就是将属性结点作为证据结点引入到贝叶斯网络中，求得类结点各取值后验概率的过程。后验概率最大时，类别结点相应的取值，即作为分类的结果，如图 2.8 所示。

在通用贝叶斯网络中，把某结点的父结点、子结点及子结点的父结点称为该结点的马尔可夫覆盖。根据有向马尔可夫性质，某结点各取值的概率只受其马尔可夫覆盖结点的影响，而与其余结点无关。

3）增强型朴素贝叶斯网络

增强型朴素贝叶斯网络(Tree-Augmented Naïve Bayes，TAN)是朴素贝叶斯网络进行有效改进的分类器，它既有朴素贝叶斯分类器的简单性，又有比朴素贝叶斯分类器更好的分类性能。TAN 是由 Friedman 提出的一种树状结构模型，它是朴素贝叶斯分类器的自然扩展，如图 2.9 所示。

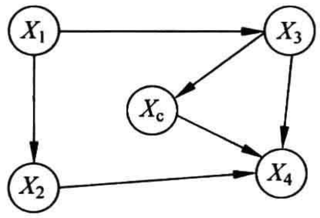


图 2.8 通用贝叶斯网络

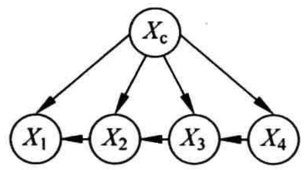


图 2.9 增强型朴素贝叶斯网络

其基本思想是将贝叶斯网络的某些表示依赖关系的能力与朴素贝叶斯的简易性相结合，使分类性能增强。

令 $U=\{X_1, X_2, \cdots, X_n, X_c\}$，其中变量 $X_1, X_2, \cdots, X_n$ 是属性变量，X_c 是类变量。在 TAN 结构中，类变量是根，没有父结点，即 $\mathrm{Pa}(X_c)=\Phi$($\mathrm{Pa}(X_c)$表示 X_c的父结点集)，类变量是每个属性变量的父结点，即 $X_c \in \mathrm{Pa}(X_i)$ [$\mathrm{Pa}(X_i)$($i=1,2,\cdots,n$)表示 X_i 的父结点集]。

属性变量 X_i 除了类变量 X_c 作为其父结点外，最多有一个其他属性变量作为其父结点，即 $\mathrm{Pa}(X_i)=2$。

因此，确定 TAN 分类模型的结构关键是如何确定每个属性结点的非类父结点，确定属性结点的父结点需要学习算法来完成。目前，标准 TAN 分类器的构造有两种方法，一种是

由 Friedman 提出的基于分布的构造算法[61]，另一种是由 Eamon 和 Pazzani[62] 提出的基于分类的构造算法。

4）马尔可夫毯贝叶斯网络

马尔可夫毯贝叶斯网络（Markov Blanket Bayesian Networks，MBBN）是由 M. G. Madden 于 2002 年提出的一种贝叶斯网络分类算法。它能表示出与通用贝叶斯网络分类器中相同的关于类结点的完整马尔可夫毯，尽管通用贝叶斯网络可以被用来进行分类任务，但是由于类结点不是被特殊看待，在类结点的马尔可夫毯以外的其他结点网络与分类任务并不相关，因而整个网络结构进行分类预测准确性不高。另外，3 个简单的贝叶斯网络，即朴素贝叶斯、增强型朴素贝叶斯，GBN，也可以用来进行分类预测，但是这些结构中都假设分类变量是根结点，忽略了类结点的父结点。然而 MBBN 并没有这种假设，比这 3 种贝叶斯网络分类器有更加丰富的结构表达，这便于更加准确地进行因果分析和分类预测。Madden 在文献[63]中比较了朴素贝叶斯、TAN、GBN 和 MBBN 这 4 种贝叶斯分类器的分类性能，结果显示 MBBN 在预测速度和准确性方面比其他 3 种分类器有相当或更好的效果，MBBN 如图 2.10 所示。

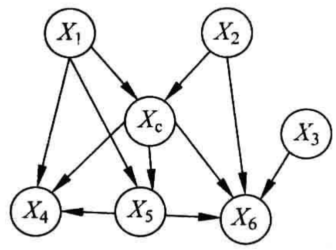


图 2.10 马尔可夫毯贝叶斯网络

MBBN 的结构学习算法如下：

步骤 1 在除 x_c 的结点中找 x_c 的父结点集 z_p，孩子结点集 z_c 和不相关结点集 z_n。当父结点因子 δ_p 大于孩子结点因子 δ_c 时，加结点 x_i 到 z_p 否则到 z_c；当 $\max(\delta_p,\delta_c)<1$ 时，加 x_i 到 z_n。

其中：

$$\delta_p = \frac{g(x_c,\pi_c \cup \{x_i\})}{g(x_c,\pi_c)} \tag{2.66}$$

$$\delta_c = \frac{g(x_i,\pi_i \cup \{x_c\})}{g(x_i,\pi_i)} \tag{2.67}$$

$$g(x_i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i} N_{ijk}! \tag{2.68}$$

π_i 和 π_c 分别表示结点 x_i 和 x_c 的父结点集；r_i 表示的 x_i 取值数目；N_{ijk} 是第 i 个结点，取值为第 k 个值时，在父结点取值组合取第 j 个值时数据的数目。

步骤 2 使用 K2 算法（见 2.2.2 小节），在除 x_c 和 z_c 的结点中找 z_c 结点的父结点。

步骤 3 使用 TAN 算法，在 z_c 中找每个结点的另一个父结点，因为最多只能找一个父结点，所以这一步所得的结构是近似的网络结构，更具体的算法说明请见文献[63]。

2.2.3 动态贝叶斯网络

贝叶斯网络反映的是事物的静态特性，而现实生活中，存在着很多的动态随机过程，动态贝叶斯网络（Dynamic Bayesian Networks，DBNs）是用来对这些过程进行建模的方法之一。

这里的“动态”表明了建模的对象是一个动态的系统，而不是系统的结构随时间发生变化。由于动态贝叶斯网络为有向图，并且对于每一个结点参数的估计可以相对独立地进行，

所以其易于推导和学习，并在近些年受到了广泛的关注。

DBNs是针对动态序列建模问题提出的。一个典型的DBNs如图2.11所示。从图中可以看出，DBNs服从马尔可夫特性：t时刻系统的所有变量的概率分布只与$t-1$时刻系统的状态变量概率分布相关。

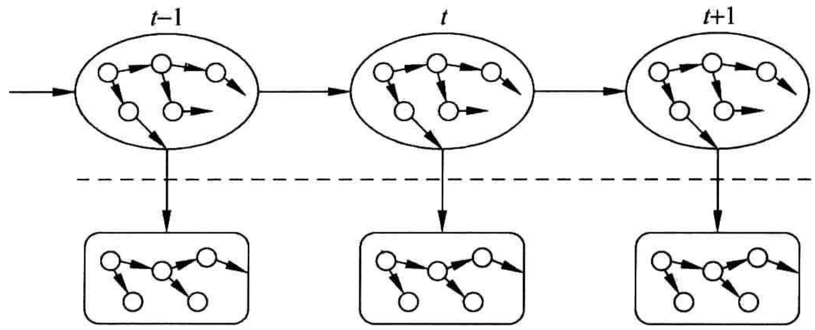
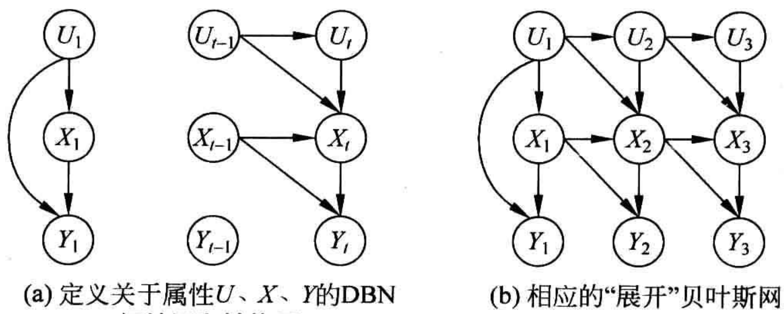


图2.11 DBNs的图模型表示

设$Z_t=(U_t,X_t,Y_t)$代表一个状态空间模型的所有变量集合，U_t代表输入变量集，X_t代表隐状态变量集，Y_t代表输出变量集，系统是离散时间随机过程，那么一个动态贝叶斯网络由一个初始网B_1和转换网$B_{\rightarrow}$组成，即$(B_1,B_{\rightarrow})$。图2.12给出了一个动态贝叶斯网络的简单例子。初始网B_1是一个贝叶斯网络，它指定了随机过程的初始条件概率分布$P(Z_1)$。转换网$B_{\rightarrow}$是含有两个时间片、由两个贝叶斯网络组成，它对所有时间点$1,2,\cdots,t$指定从时间点$t-1$到时间点t属性集状态的转换概率$P(Z_t\mid Z_{t-1})$。根据有向无环图(DAG)[见图2.12(b)]，得下面公式，即

$$p(Z_t \mid Z_{t-1}) = \prod_{i=1}^{N} p(Z_t^i \mid \mathrm{Pa}(Z_t^i)) \tag{2.69}$$

式中，Z_t^i为时刻t的贝叶斯网中第i个结点，即U_t、X_t或者Y_t其中之一的一个结点；$\mathrm{Pa}(Z_t^i)$为Z_t^i的父结点集的取值；第一个时间片BN中的结点之间没有任何相关参数，第二个时间片的BN中结点之间存在相关的条件概率分布CPD，这定义了观测条件概率分布$p(Z_t^i\mid\mathrm{Pa}(Z_t^i))(t>1)$，这些CPD的形式是任意的；这里假设条件概率分布的参数是不随时间变化的，并且转移概率$P(Z_t\mid Z_{t-1})$也是时不变的。



(a) 定义关于属性U、X、Y的DBN初始网和转换网　　(b) 相应的“展开”贝叶斯网

图2.12 贝叶斯网及时间片的展开

一个动态贝叶斯网络定义了在动态随机过程中无穷变化轨迹上的概率分布。实际上，一般只在有穷时间间隔$1,2,\cdots,T$上推理，那么可以把一个动态贝叶斯网络展开成在Z_1，$Z_2,\cdots,Z_T$上的“长”贝叶斯网络。图2.12(b)给出了图2.12(a)所示的动态贝叶斯网展开3

个时间片的相应贝叶斯网。给定动态贝叶斯网络 $B=(B_1,B_{\rightarrow})$，在 $Z_1,Z_2,\cdots,Z_T$ 上的联合概率分布可以通过初始网和转换网指定的概率分布简化表示为

$$p(Z_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{N} p(Z_t^i \mid \mathrm{Pa}(Z_t^i)) \tag{2.70}$$

完全确定 DBNs 需要知道 3 个概率分布：状态转移条件分布 $P(Z_t|Z_{t-1})$、观测条件分布 $P(Z_t^i|\mathrm{Pa}(Z_t^i))$ 和初始状态分布 $P(Z_1)$。所有的条件分布可以是时变的或定常的，可以取参数化形式 $P(x_t|x_{t-1};\theta)$，也可以使用非参数化(概率表格或统计直方图)表示。

隐马尔可夫模型和卡尔曼滤波模型就是两种典型的状态空间模型。因为它们简单和灵活而成为研究随机过程的主要方法。例如，隐马尔可夫模型已经被应用在语音识别和生物序列分析中；卡尔曼滤波模型应用于跟踪飞机和导弹的轨迹及经济预测等。然而 HMMs 和 KFMs 都受限于其表达知识的能力上，然而动态贝叶斯网络作为带有时间参数的贝叶斯网络可以克服其这个缺点。动态贝叶斯网络能够用一个有 N_h 个变量的集合来表达隐状态，$X_t^1,\cdots,X_t^{N_h}$，即分布式的状态表达；相反，隐马尔可夫模型隐状态的表达是一个有 M 个可能的离散变量 X_t。动态贝叶斯网络允许条件概率分布是任意的，而卡尔曼滤波模型要求条件概率分布必须是线性-高斯分布。另外，隐马尔可夫模型和卡尔曼滤波模型的拓扑结构有限，而动态贝叶斯网络允许有更多样的图形结构。实际上，隐马尔可夫模型和卡尔曼滤波模型可以作为动态贝叶斯网络的特例。

2.3 无向概率图模型

无向概率图模型 (Undirected Probabilistic Graphical Model)使用一个无向图建立随机变量之间的关系模型。无向链接通常捕捉一对结点之间的相互依赖关系。马尔可夫随机场(Markov Random Fields,MRFs)和条件随机场(Conditional Random Fields,CRFs)是两种无向概率图模型，其被广泛应用在图像处理、目标识别、图像分割和纹理合成等计算机视觉领域。

2.3.1 马尔可夫随机场

马尔可夫随机场(MRF)，也叫做马尔可夫网，由于其对变量空间关系及上下文独立关系建模的灵活方便的特性，受到了计算机视觉研究领域的高度关注。图像分析是马尔可夫随机场应用最广泛的领域，一些重要的研究工作包括 1986 年 J. Besag 发表的 *On the statistical analysis of dirty pictures* 以及 2004 年 S. Geman 在 PAMI 期刊上发表的 *Stochastic relaxation, gibbs distributions, and bayesian restoration of images* 文章。

马尔可夫随机场是关于一组有马尔可夫性质随机变量 X 的全联合概率分布模型。马尔可夫网络类似贝叶斯网络用于表示依赖关系。但是，一方面，它可以表示贝叶斯网络无法表示的一些依赖关系，如循环依赖；另一方面，它不能表示贝叶斯网络能够表示的某些关系，如推导关系。马尔可夫网络的原型是 Ising 模型，最初是用来说明该模型的基本假设。

1. 条件独立性

在马尔可夫网络中，关于独立的性质有这样的结论：如果 X_B 把 X_A 和 X_C“分割”开，则

给定 X_B 时，X_A 与 X_C 相互独立，这样的性质也称为马尔可夫性。"分割"开是指的图论意义上的"分割"开，如图 2.13 所示（X_B 是图的一个点割集）。也可以这样表达，如果从 X_A 中的一个结点到 X_C 中的一个结点的任意一条路径都至少包含有 X_B 中的一个结点，则可以断言给定 X_B 时，X_A 和 X_C 相互独立；否则，给定 X_B 时，X_A 和 X_C 不相互独立。有了独立性定义，关于独立性检验的算法就变得非常直接：只需把 X_B 从中移走，X_A 和 X_C 是否相互独立就可以通过考察是否存在从 X_A 到 X_C 的路径。这在图论中被称为"连通性"问题（Conectivity）。图论中提供了标准的搜索算法。

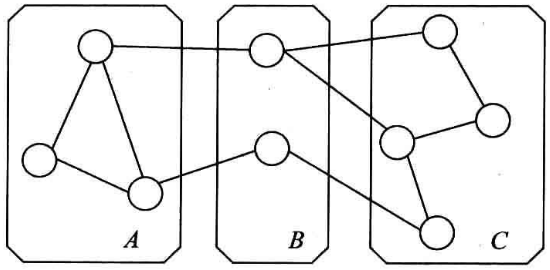


图 2.13 结点集 X_B 将 X_A 和 X_C"分割"开

2. 联合概率分布

根据 Hammersley-Clifford 定理和局部马尔可夫性质，无向图的概率分布可以被定义为下面 Gibbs 公式，即

$$P(x) = \frac{1}{Z}\prod_{C}\psi_C(x_C) \tag{2.71}$$

其中，$\psi_C(x_C)$是一个关于 x_C 的非负实数函数（实值函数），通常被称为团的势函数，变量 x_C 属于集合 C。Z 是归一化常数（或称为 Partition Function），其取值为 $Z \equiv \sum_x \prod_C \psi_C(x_C)$。在马尔可夫随机场中，不像贝叶斯网那样，作为局部条件和边缘分布，每个归一化常数中的势函数 $\psi_C(x_C)$没有方向的意义。这一方面给用语义定义局部函数带来了更多的灵活性；但另一方面，它又包括了计算归一化常数及其派生函数，这使得模型学习和推理变得困难。图 2.14(a)示出了一个用马尔可夫随机场为图像的空间关系所建立的图模型。

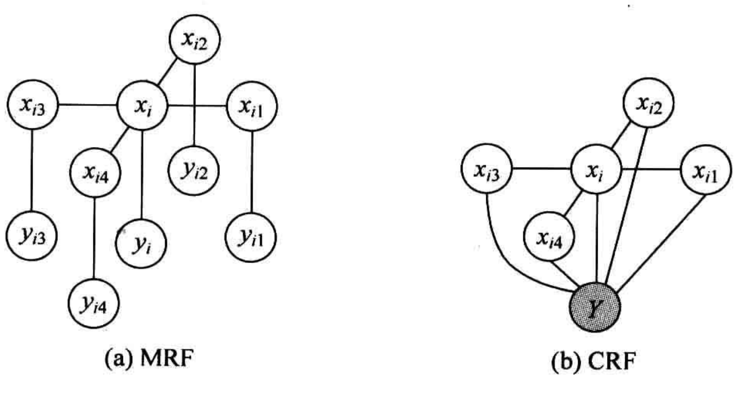


图 2.14 无向图模型

在计算机视觉领域中研究和应用马尔可夫网络，人们感兴趣的是如何给出马尔可夫网络中变量的联合概率分布。另外，对非规则结构的马尔可夫网络模型的研究主要集中在统计领域，如对数—线性模型和专家系统。在这些研究中，人们更为关注的是如何在一个给定

的马尔可夫网络模型中进行概率推理的问题。非规则结构的马尔可夫网络模型在应用中常常存在两个问题,一是指定网络中所有变量的联合概率分布往往是困难的,因为其中的参数缺少明确的物理意义,二是马尔可夫网络模型刻画的是随机变量之间对称的相关关系,而在许多应用中变量之间的相互作用并不具有对称性。例如,在专家系统中,变量之间往往具有较为明确的因果关系,那么如何描述这种因果关系,或者说如何利用此因果关系来建立恰当的模型?对这种因果性、时序性的描述,使用贝叶斯网络常常是适合的。

2.3.2 条件随机场

最近,在语音处理、信息提取和计算机视觉领域,对结构化预测问题建立判别模型受到了广泛的关注。条件随机场(CRF)就是这些最成功的判别模型之一,它在图像分割、形状分析和图像标注等方面优于生成模型,如马尔可夫随机场。条件随机场是一种无向图模型,当给定观测值(用 y 表示),其能够直接表达类条件概率分布(类变量用 x 表示),即

$$P(x \mid y) = \frac{1}{Z(y)} \prod_{C} \psi_C(x_C, y) \tag{2.72}$$

其中,$\psi_C(x_C, y)$不仅依赖于集合 x,同时也依赖于整个观测值 y。图 2.14(b)示出了用 CRF 对图像空间关系建立的图模型。

CRFs 同 MRFs 相比,最主要的优点有:CRFs 关注于最终的预测问题,避免不必要的观测密度计算;CRFs 不要求像生成模型中对于观测变量之间条件独立关系的假设。并且条件模型的优势(如较低的渐进预测误差)已经在理论上与实践中得到了验证。

2.4 概率图模型学习与推理

与概率图模型相关的两个主要问题是学习和推理(Leaning and Inference)。模型的学习是从数据中学习模型的结构和参数,而推理的目的是对给定的模型计算边缘分布和条件分布。模型的学习和推理都涉及算法的计算复杂度,一般来讲精确的学习和推理都是 NP-Hard 问题。

2.4.1 模型的学习

概率图模型的学习包括结构学习和参数学习两个方面。

1. 结构学习

结构学习(Structure Learning) 也被称为模型选择,可分为两类:一类是基于评分函数(Scoring Function)的学习;另一类是基于独立性测试(Conditional Independence Test, CIT)的学习。基于评分函数的学习是用一个预定义好的函数作为评分标准,对模型结构空间中所有的模型进行评分,选出分值最大者所对应的网络结构作为最终网络结构。常用的评分函数有基于贝叶斯统计的 BDe(Bayesian Dirichlet-Likelihood Equivalent)、最小描述长度 MDL(Minimum Description Length)和贝叶斯信息标准 BIC(Bayesian Information Criterion)。当然,在巨大的模型空间作盲目搜索,要得到最终的网络结构是相当困难的,故而常采用贪心策略、模拟退火、最优最先等算法进行搜索。基于独立性测试的学习是通过度

量属性之间的独立性关系,确定模型结构的方法,常用的算法有 CL 算法、3 阶段算法等。基于评分函数的学习是一种常用的方法,其中 BIC 评分函数和 BDe 评分函数最为常使用。

2. 参数学习

给定数据集的概率图模型参数学习(Parameter Learning)可以被看作是最大化数据似然的优化任务。优化任务的复杂性往往根据不同的因素而变化,如数据是完全观测还是部分观测、是有向图还是无向图、推理是否在学习过程中使用等。

在完全观测数据的情况下,学习一个有向贝叶斯网络是非常容易的,因为它要求父-子结点是条件独立,这经常会导致一个闭型解。然而,学习一个无向马尔可夫模型是非常复杂的,因为计算分割函数存在一定的困难。

当数据不完整或存在隐变量时,EM(Expectation-Maximization)算法经常被使用,它包含两个步骤:第一步是 E-step,利用当前网络结构和参数对缺失数据或隐变量计算它的期望值;第二步是 M-step,基于 E-step 参数的期望值,计算出新的最大可能的参数分布 θ',用 θ' 替换原有的 θ。

EM 算法也被广泛应用到不完整数据的结构学习中。结构 EM 算法,其中包含相同的 E-step,在它的 M-step,基于当前的完整样本集,进行模型选择,即评价任意一个候选网络结构,其网络结构的评价函数可以是包括 BIC 评分在内的任意一个评分函数,那么最终的结构应该是分值最高的。

2.4.2 模型的推理

设一个由 n 个变量构成的联合概率分布表达了该 n 个变量各种组合的概率值,若每个随机变量有两个值,则联合概率分布将会包含 2^n 个概率值,要计算各种概率值,就可以通过边缘化(Marginalization)算子进行边缘概率的求取。例如,求边缘概率 $P(Y_1)$ 可以按式(2.73)计算,即

$$P(Y_1) = \sum P(Y_1, Y_2, \cdots, Y_n) \tag{2.73}$$

其中 $\sum$ 表示对集合 Y 中去掉变量 Y_1 后对所有变量求和。

以上的计算概率的方法虽然简单,但需要完成不符合实际的指数式求和运算量,人们提出了很多简化的方法。虽然 Cooper 在 1990 年证明在任意贝叶斯网上的精确推理计算是 NP-Hard 的,但是在实际问题中往往能够根据不同的贝叶斯网结构寻找到一种方式对联合概率分布进行参数化,然后在参数化的基础上寻找局部化的计算过程,以加快推理计算速度。而它们之间的不同之处在于每种方法参数化的方式不同或寻找局部计算的途径不同。

1. 精确推理

(1) 基于网络结构的推理

该方法是利用贝叶斯网络的条件独立性进行推理的,它利用了网络的结构信息,令联合概率实现了局部化。下边以一个例子来加以说明,如图 2.15 所示。

设要计算出条件概率 $P(f|a,s,g,j)$,可以由贝叶斯公式得

$$P(f \mid a,s,g,j) = \frac{P(f,a,s,g,j)}{P(a,s,g,j)} = \frac{P(f,a,s,g,j)}{\sum_{f} P(f,a,s,g,j)} \tag{2.74}$$

由链规则及条件独立性有

$P(f,a,s,g,j)=P(f)P(a)P(s)P(g|f)P(j|f,a,s)$ 则

$$P(f \mid a,s,g,j) = \frac{P(f)P(a)P(s)P(g \mid f)P(j \mid f,a,s)}{\sum_{f} P(f)P(a)P(s)P(g \mid f)P(j \mid f,a,s)} \tag{2.75}$$

这样通过各结点的条件概率就可以完成推理工作。该方法简单易行，在网络为稀疏网络时，还是较为有效的。

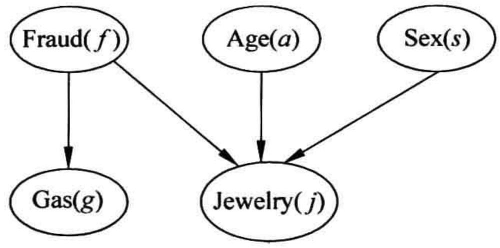


图 2.15　一个用于推理的贝叶斯网例子

(2) 精确推理的其他算法

现在已经提出了多种精确推理的方法，常用的有图形简化法、证据扩散与汇聚算法、变量消元法、联合树算法等。图形简化法是在贝叶斯网图形上的结点消元来模拟边缘概率的计算，包含结点删除(Node Removal)和弧反向(Arc Reversal)两个过程；证据扩散与汇聚算法适用于单连通结构的贝叶斯网络，对于多连通结构，由于两结点间存在多条路径，可能会使证据传递在两个结点间往返多次；变量消元法首先根据条件独立将联合概率分解为多个因子项，然后在符号层面上对公式进行变换，改变求和时结点的消元顺序以及求和运算与乘积运算的先后顺序，以达到减少求和及乘积运算量的目的，最后按变换后的公式进行逐步的乘积及求和运算以得到待求结果；联合树算法在无证据的条件下，变量求取的主要思路是将贝叶斯网络经过图形变换变成联合树，经初始化后，即将贝叶斯网络的条件概率表转化到联合树中，得到转化成非全局一致的联合树，通过消息传递，使之达到全局一致，最后通过边际化得到变量结点的分布。当有证据时，流程与无证据时类似，只是在进行初始化时，需要加入观测值，以形成非全局一致的联合树，消息传递生成一致的联合树后，通过边际化和正规化得到条件概率分布。

2. 近似推理

在任意结构的贝叶斯网络上的精确推理都是NP-Hard的。目前提出了多种近似推理方法，这些算法采取某种方式，在运行时间和推理精度上寻求一个折衷，力求在短时间内给出一个满足精度要求的结果。贝叶斯网络的近似推理算法主要分为两类：一类是基于仿真的方法，通过使用一个包含随机数发生器的采样装置，来根据需要产生一组样本，然后通过对样本的处理，以求得待求概率的近似值，这类方法都是基于Monte Carlo的基本思想；另一类方法是基于搜索的方法，该方法认为概率问题是一类组合问题，可以将所需要计算的各个变量的不同组合看作一个状态空间。这个空间中存在一些状态，它们对最后的计算结果

会产生较大影响，而另外一些状态则影响不大。由此可以通过启发式搜索方法，在整个状态空间中进行搜索，以寻找到这些状态。从而以这些状态代替整个状态空间参与运算，达到提高计算效率的目的，并能得到较精确的解答。

2.4.3 计算复杂度分析

对于简单的概率图模型（如那些无环图），精确推理是十分有效的。然而对于复杂的概率图模型，精确推理（如变量消元算法、信度传播算法或者消息传递算法）被看作是联合树推理算法的特例。

联合树算法需要三角化过程，其实质上是找到清除结点连线的顺序。找到优化的三角化过程被认为是 NP-Hard 问题。通常，在牺牲一定的准确性下，用可实现的近似推理算法来解决精确推理遇到的困难，或者设定可替代的伪似然学习目标。然而，基于近似抽样的推理方法经常出现慢收敛。

在完整观测数据下，贝叶斯网络的参数学习几乎是不可实现的，然而用部分观测数据学习贝叶斯网络或者是马尔可夫网络要求每一迭代过程中都要进行推理，这被认为是 NP-Hard 问题。并且，结构学习比参数学习还要复杂得多。一些理论研究已经显示，即使学习一个有界结构的模型，都是 NP-Hard 问题（例如，一个有界树宽的马尔可夫网络，一个有限父结点数目的贝叶斯网络和 Poly 树）。然而，最近几种多项式时间的算法已经被提出来，这些算法为马尔可夫网络和因子图提供了较好的近似计算[312,313]。

第3章 完整数据集下有向概率图模型的学习

3.1 概述

概率图模型建模一般有 3 种方法：依靠专家建模；从数据中学习；从知识库中创建。实际建模过程中常常综合运用这些方法，以专家知识为主导，以数据库和知识库为辅助手段，扬长避短，发挥各自优势，来保证建模的效率和准确性。但是，在不具备专家知识或知识库的前提下，从数据中学习概率图模型显得尤为重要。

本章介绍了在完整数据集下有向概率图模型的学习方法。完整数据集意味着数据集中的各样本属性特征数据完整，样本数据独立分布。有向概率图模型主要以贝叶斯网络为学习对象，分别从结构学习和参数学习两个方面介绍常用的学习方法。

在结构学习方面常用的方法有两种：一种是基于评分函数（Scoring Function）的学习；另一种是基于独立性测试（Conditional Independence Test，CIT）的学习。

在参数学习方面，最常用的方法就是极大似然参数估计和贝叶斯参数估计。

3.2 结构学习

3.2.1 基于评分-搜索的结构学习

基于评分-搜索的结构学习（Structure Scores）主要由两部分组成，即评分函数和搜索算法。常用的评分函数包括 BDe 评分函数、MDL 评分函数和 BIC 评分函数。搜索算法包括启发式局部搜索算法和全局搜索算法。

1. 评分函数

（1）BDe 评分函数

基于 BDe（Bayesian Dirichlet-Likelihood Equivalent）评分函数（BDe Score Function）的结构学习是以贝叶斯统计学作为理论基础的，它的主要思想是：在假设参数分布的条件下，计算某一结构相对于给定数据集的后验分布，将该分布作为选择结构的依据，找出后验分布最大的结构。

对于全部由离散变量构成的贝叶斯网络的学习，常常是将其转化为一系列多态抽样过程来进行学习的。在很多算法的研究中，一般选用多态分布的共轭分布——Dirichlet 分布作为参数 θ 的先验分布。

Dirichlet 分布的公式为

$$\mathrm{Dir}(\theta \mid \alpha_1,\alpha_2,\cdots,\alpha_r)=\frac{\Gamma\left(\sum_{k=1}^{r}\alpha_k\right)}{\prod_{k=1}^{r}\Gamma(\alpha_k)}\prod_{k=1}^{r}\theta_k^{\alpha_k-1}\quad \alpha_k>0 \tag{3.1}$$

式中，$\alpha_1,\alpha_2,\cdots,\alpha_r$ 为超参数；$\Gamma(\cdot)$为伽马函数，定义为

$$\Gamma(n)=\int_0^{\infty}t^{n-1}\mathrm{e}^{-t}\mathrm{d}t\quad n>0 \tag{3.2}$$

假设数据集变量均为离散值，B_G 表示某一网络结构，由贝叶斯公式可以得到

$$P(B_G \mid D)=\frac{P(D \mid B_G)P(B_G)}{P(D)} \tag{3.3}$$

由于数据集的先验概率 $P(D)$都相等，故在对各网络结构进行比较时，可以不予考虑；$P(B_G)$是网络结构的先验概率，可以将它们视为相同，所以要计算 $P(B_G|D)$的值，主要需要计算 $P(D|B_G)$，在无隐含结点的条件下，有

$$P(D \mid B_G)=\int_{\boldsymbol{B}_P}P(D \mid B_G,\boldsymbol{B}_P)f(\boldsymbol{B}_P \mid B_G)\mathrm{d}\boldsymbol{B}_P \tag{3.4}$$

式中，$\boldsymbol{B}_P$ 为相对 B_G 的条件概率分布向量；$f(\cdot)$是已知 B_G 条件下 $\boldsymbol{B}_P$ 的条件概率分布函数。若数据集中含有的 m 个实例 X_k，且他们之间独立，式(3.4)可以变为

$$P(D \mid B_G)=\int_{\boldsymbol{B}_P}\left[\prod_{k=1}^{m}P(X_k \mid B_G,\boldsymbol{B}_P)\right]f(\boldsymbol{B}_P \mid B_G)\mathrm{d}\boldsymbol{B}_P \tag{3.5}$$

在数据集完备并且各参数 $\vec{\theta}_{ij}$ 的先验分布相互独立的条件下，有

$$P(D \mid B_G)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\int_{\theta_{ijk}}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}}f(\theta_{ij1}\theta_{ij2}\cdots\theta_{ijr_i})\mathrm{d}\theta_{ij1}\cdots\mathrm{d}\theta_{ijr_i} \tag{3.6}$$

式中，θ_{ijk}为在 X_i 取第 k 个值，其父结点取第 j 个组合值时的概率；α_{ijk}是对应于 θ_{ijk}的实例数目。

假设 $f(\theta_{ij1}\theta_{ij2}\cdots\theta_{ijr_i})$服从 Dirichlet 分布，则可以得到 BDe 评分函数计算公式为

$$P(D|B_G)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha'_{ij})}{\Gamma(\alpha'_{ij}+\alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha'_{ijk}+\alpha_{ijk})}{\Gamma(\alpha'_{ijk})} \tag{3.7}$$

这里 α'_{ijk}是超参数，并且 $\alpha'_{ij}=\sum_{k=1}^{r_i}\alpha'_{ijk}$，$\alpha_{ij}=\sum_{k=1}^{r_i}\alpha_{ijk}$，$\Gamma(\cdot)$是伽马函数，定义如前所述。

对于超参数 α'_{ijk}的估计是非线性的，在计算上有很大困难。G. Cooper 和 E. Herskovits 提出可以采用 $\alpha'_{ijk}=1$ 来进行计算；当然还可以用其他方法来估计，这里不再赘述。

(2) MDL 评分函数

最小描述长度 MDL(Minimum Description Length)是 Rissanen 在研究通用编码的时候提出来的。假如给定数据集 D，如果要对其进行保存，为节省存储空间，可以采用某种模型对其进行编码压缩，然后再保存压缩后的数据。并且为了能够恢复 D，还必须存储这个压缩模型。因此，需要保存的数据长度等于这些数据进行编码压缩后的长度加上保存模型所需的长度，这个长度就被定义为总描述长度。MDL 原理就是要选择总描述长度最小的模型。

在数据完备的条件下，总的描述长度的值为

$$\mathrm{DL}(G,D) \approx \sum_{i=1}^{n} \mid \mathrm{Pa}_i \mid \log_2(n) + \frac{1}{2}\sum_{i=1}^{n} \mid \mathrm{Pa}_i \mid (\mid X_i \mid -1)\log M + M\sum_{i=1}^{n} H(X_i \mid \mathrm{Pa}_i) \tag{3.8}$$

式中，$H(X_i|\mathrm{Pa}_i)$为 X_i 对于 Pa_i 的条件熵。

MDL 的思想就是最小化 $\mathrm{DL}(G,D)$，等价于最大化$(-\mathrm{DL}(G,D))$，并且只保留了随 M 增长的因子，故式(3.8)可以进一步简化为

$$-\mathrm{DL}(G,D) \approx -M\sum_{i=1}^{n} H(X_i \mid \mathrm{Pa}_i) - \frac{1}{2}\sum_{i=1}^{n} \mid \mathrm{Pa}_i \mid (\mid X_i \mid -1)\log M \tag{3.9}$$

式中，M 为数据集长度。

MDL 评分的计算过程中没有用到参数的先验值，不需要对参数进行估计，在这一点上，它的计算比 BDe 评分的计算简单。

(3) BIC 评分函数

BIC(Bayesian Information Criterion)评分实质上是对 BDe 评分的一种近似计算。

在实际使用 BDe 评分时，一般不是直接计算 $P(B_G|D)$，而是求它的对数 $\log P(B_G|D)$，BIC 评分的计算公式为

$$\log P(B_G \mid D) = \log P(D \mid \tilde{\theta}_G, B_G) - \frac{d}{2}\log M \tag{3.10}$$

式中，d 为高斯分布函数的维数，即其中的变量数目，对于贝叶斯网有 $d=\sum_{i=1}^{n}|\mathrm{Pa}_i|(|X_i|-1)$；$M$ 为数据集的数目；$\tilde{\theta}_G$ 为 $P(\theta_G|D,B_G)$后验分布最大时 θ_G 参数的值。

BIC 评分是 BDe 评分在数据集数据数目 M 很大时的近似计算。因此，当数据集数据数目很大时，这 3 个评分函数的计算结果趋向一致。

BDe 评分在学习时需要考虑参数的先验分布，而在基于 MDL 和 BIC 评分的结构学习中，都是采用极大似然估计法估计出该结构的参数值。这些值可以通过对数据集的学习自动获得，不需要人工干预。所以计算 MDL 和 BIC 评分要比计算 BDe 评分简单。然而，计算的简单是以忽略先验知识为代价的，由于 MDL 和 BIC 评分没有用到先验知识，采用这两种评分学习结果的正确性完全依赖于实例数据集合，这就要求实例数据必须很大，并且不能出现大的偏差。而使用 BDe 评分学习时，尽管其结构的正确性也依赖于实例数据集合，但通过对参数的先验值的估计，其依赖性比 MDL 与 BIC 评分要低，有时甚至可以纠正实例数据中的偏差。

从学习结构复杂性的角度来看，BDe 评分中并没有明确地包含结构复杂性指标，在某些情况下，该评分会倾向于选择较为复杂的网络结构；而在 MDL 和 BIC 评分中，明确地将结构复杂性作为一个指标，因此，这两个评分倾向于选择较简单的网络结构，在实际学习中，采用这两个评分计算的结果更简单、更容易被接受。

2. 搜索算法

由于贝叶斯网络的结构学习问题是一个 NP-Hard 问题，所以在实际计算中，并不是对所有的结构分别计算其评分值，再进行比较取最优，而是采用搜索算法(Structure Search)，按照某种评分在可能的拓扑结构空间中进行搜索来获取结构。

最基本的搜索算法是启发式局部搜索算法，主要有 K2 算法，Hill-climbing 算法、随机重复爬山法(Random Restart Hill-climbing)、禁忌搜索(Tabu Search)、模拟退火(Simulated Annealing)及遗传算法(Genetic Algorithm)等。搜索过程主要包括 3 个阶段：

阶段 1(Drafting)：计算每对结点间的互信息，建立完整的无向图。

阶段 2 (Thickening)：如果结点对不是 *d*-分割的话，把这一点对加入到边集中。

阶段 3(Thinning)：检查边集中的每个点对，如果两个结点是 *d*-分割的，则移走这条边。

典型算法分析如下：

(1) K2 算法

K2 算法用贪婪搜索处理模型选择问题：先定义一种评价网络结构优劣的评分函数，再从一个网络开始，根据事先确定的最大父结点数目和结点次序，选择分值最高的结点作为该结点的父结点。K2 算法使用 BDe 评分函数。

K2 算法伪代码如下：

Algorithm3.1 K2 algorithm for BN structure learning

K2(

$X=\{X_1, X_2, \cdots, X_n\}$, //一组变量

ρ, //一个变量顺序(设它与变量下标一致)

N, //变量父亲结点的最大个数

D, //一组完整的数据

)

1 $\zeta\leftarrow$由结点 $X_1, X_2, \cdots, X_n$ 组成的无边图

2 For j=1 to n

3 $\pi_j\leftarrow\phi$;

4 $V_{old}\leftarrow$BDe($<X_j, \pi_j>|D$)

5 While(true)

6 $i\leftarrow \arg\max_{1\leq i<j, X_i\notin\pi_i}$ BDe($<X_j, \pi_j\cup\{X_i\}>|D$)

7 $V_{new}\leftarrow$BDe($<X_j, \pi_j\cup\{X_i\}>|D$)

8 $If(V_{new}>V_{old}$ and $|\pi_j|<N)$

9 $V_{old}\leftarrow V_{new}$;

10 $\pi_j\leftarrow\pi_i\cup\{X_i\}$;

11 在 D 中加边 $X_i\rightarrow X_j$;

12 Else

13 Break;

14 End if

15 End While

16 End for

17 估计 ζ 的参数 θ

18 Return (ζ, θ); //输出一个贝叶斯网络

K2 的出发点是一个包含所有结点，但却没有边的无向图。在搜索过程中，K2 按顺序逐个考察 ρ 中的变量，确定其父结点，然后添加相应的边。对某一变量 X_j，假设 K2 已经找到了它的一些父结点 π_j。如果 $|\pi_j|<N$，即 X_j 的父结点个数还未达到上界 N，那么就要继

续为它寻找父结点，具体做法是首先考虑那些在 ρ 中排在 X_j 之前，但却还不是 X_j 的父结点的变量，从这些变量中选出 X_i，它使得新家族 BDe 评分 $V_{new} \leftarrow \text{BDe}(<X_j, \pi_j \cup \{X_i\}>|D)$ 达到最大；然后将 V_{new} 与旧家族评分比较：如果 $V_{new} > V_{old}$，则把 X_i 添加为 X_j 的父结点；否则停止为 X_j 寻找父结点。

(2) Hill-climbing 算法

此法的目标是要找出评分最高的模型，它从一个初始模型出发开始搜索，初始模型一般设为无边模型，在搜索的每一步，它首先用搜索算子对当前模型进行局部修改，得到一系列候选模型；然后计算每个候选模型的评分，并将最优候选模型与当前模型进行比较；若最优候选模型的评分大，则以它为下一个模型继续搜索；否则，就停止搜索，并返回当前模型。

搜索算子有 3 个：加边、减边和转边。加边和减边算子的使用有个前提，就是不能在网络中形成有向圈。

此法可以使用任何评分函数。不同的评分函数有不同的要求：BDe 评分要求关于先验参数分布的超参数；HVL(Holdout Validation Likelihood，验证数据似然度)评分要求及 CVL(Cross Validation Likelihood，交叉验证)评分要求则是把数据分成训练数据和验证数据。因此，需要处理的算法细节也有所不同。

Hill-climbing 算法的伪代码如下：

Algorithm3.2 Hill Climbing algorithm for BN structure learning

```
LearnBN_HC(
X,  //一组变量
ϑ,  //一组关于 X 的完整数据
f,  //一个罚项似然度评分函数
ζ0, //一个初始贝叶斯网络结构
)
1  ζ←ζ0;θ←ζ 的参数的最大似然估计
2  oldScore←f(ζ,θ|ϑ);
3  While(true)
4      ζ* ←null;θ* ←null;newScore←-∞
5      For(每个对 ζ 做一次加边、减边或逆转边而得到的模型结构 ζ')
6          θ'←ζ'的参数的最大似然估计;
7          tempScore←f(ζ',θ'|ϑ);
8          If(tempScore>newScore)
9              ζ* ←ζ';θ* ←θ';newScore←tempScore;
10         End If
11     End For
12     If(newScore>oldScore)
13         ζ←ζ* ;θ←θ* ;oldScore←newScore;
14     Else
15         Return(ζ,θ);
16     End If
17 End While
```

3.2.2 基于条件独立性测试的结构学习算法

基于条件独立性测试(Conditional Independence Test)的结构学习算法的核心思想是：

首先对训练数据集进行统计测试，尤其是条件独立性测试，确定出不同结点集之间的一致条件独立性；然后，利用结点集之间的条件独立性，构造一个有向无环图，以尽可能多地涵盖这些条件独立性。基于统计测试的方法比较容易理解，这与贝叶斯网络的定义紧密相连，它将独立的概念从结构构造中分离出来。

1993年Spirtes等提出的SGS算法是典型的以条件独立性测试确定拓扑结构的算法。该算法从无向完全图出发，如果相邻结点间存在无向分割集，则删除它们之间的边；然后通过统计测试来确定剩余边的方向。

2002年，Cheng将信息论与统计测试相结合，使用相互信息代替了条件独立测试，经过Drafting、Thickening、Thinning 3个步骤，通过计算互信息量(Mutual Information)来确定结点间的条件独立性，从而构造多连接有向模型。

3.3 参数学习

在参数学习(Parameter Learning)方面，参数估计问题最常用的和有效的方法就是：极大似然参数估计和贝叶斯参数估计。极大似然把待估计的参数看作是确定性的量，只是其取值未知。最佳估计就是使得产生以观测到的样本的概率最大的那个值。贝叶斯估计则把待估计的参数看成是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，这样就利用样本的信息修正了对参数的初始估计值。在贝叶斯估计中，一个典型的效果就是，每得到新的观测样本，都使得后验概率密度函数变得更加尖锐，使其在待估计参数的真实值附近形成最大的尖峰。

3.3.1 极大似然参数估计

1. 概念

极大似然估计(Maximum Likelihood Estimation，MLE)方法是求估计的一种方法。它于1821年首先由德国数学家C. F. Gauss提出，但是这个方法通常被归功于英国的统计学家R. A. Fisher，他在1922年的论文 *On the mathematical foundations of theoretical statistics* 中再次提出了这个思想，并且首先探讨了这种方法的一些性质。极大似然估计这一名称也是Fisher命名的。这是一种目前得到广泛应用的方法。它是建立在极大似然原理基础上的一个统计方法。

2. 极大似然估计基本原理

在一次观察中，一个事件出现了，可以认为此事件发生的概率大。例如，一个随机试验如有若干个可能的结果$A,B,C,\cdots$。若在一次试验中，结果A出现，则一般认为试验条件对A出现有利，也即A出现的概率很大。

下面对连续型和离散型母体两种情形阐述极大似然估计原理(The Maximum Likelihood Principle)。

设$\xi_1,\xi_2,\cdots,\xi_n$为取自具有概率函数$\{f(x,\theta):\theta\in\Theta\}$的母体$\xi$的一个子样。子样$\xi_1,\xi_2,\cdots,\xi_n$的联合概率函数在$\xi_i$取已知观测值$X_i\,(i=1,\cdots,n)$时的值$f(x_1;\theta)f(x_2;\theta)\cdots$

$f(x_n;\theta)$是θ的函数。用$L(\theta)=L(\theta;X_1,\cdots,X_n)$表示,称为这个子样的似然函数。于是有

$$L(\theta)=L(\theta;X_1,\cdots,X_n)=f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) \tag{3.11}$$

如果母体是离散型母体,$L(\theta)=L(\theta;X_1,\cdots,X_n)$给出观测到$(X_1,X_2,\cdots,X_n)$的概率。因此,可以把$L(\theta)=L(\theta;X_1,\cdots,X_n)$看成为了观测到$(X_1,X_2,\cdots,X_n)$时出现什么样$\theta$的可能性的一个测度。所以只要寻找这样的观测值$(X_1,X_2,\cdots,X_n)$的函数$\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_n)$,以$\hat{\theta}$代替$\theta$使

$$L(\theta;X_1,\cdots,X_n)=\sup_{\theta\in\Theta}L(\theta;X_1,\cdots,X_n) \tag{3.12}$$

成立。满足式(3.12)的$\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_n)$就是最可能产生$X_1,X_2,\cdots,X_n$的参数$\theta$的值。则称$\hat{\theta}(X_1,\cdots,X_n)$为参数$\theta$的极大似然估计值,其相应的统计量$\hat{\theta}(\xi_1,\cdots,\xi_n)$称为参数$\theta$的极大似然估计量。

如果ξ是连续型,$f(x_1;\theta)$,$\theta\in\Theta$表示密度函数,Θ是参数空间。于是子样$\xi_1,\xi_2,\cdots,\xi_n$落入$(X_1,X_2,\cdots,X_n)$的邻域内的概率为$\prod_{i=1}^{n}f(x_i;\theta)\Delta x_i$,同样是$\theta$的函数。既然$(X_1,X_2,\cdots,X_n)$在一次抽样中出现,当然可以认为子样$(\xi_1,\xi_2,\cdots,\xi_n)$落在$(X_1,X_2,\cdots,X_n)$邻域内的概率达到最大。所以只要找出使$\prod_{i=1}^{n}f(x_i;\theta)\Delta x_i$达到最大的$\theta$的值$\hat{\theta}(X_1,\cdots,X_n)$。由于$\Delta x_i$是不依赖于$\theta$的增量,也只需求出使得

$$L(\theta;X_1,\cdots,X_n)=\prod_{i=1}^{n}f(x_i;\theta)\Delta x_i \tag{3.13}$$

达到最大值,便可得到极大似然估计。综上所述知道,连续母体的参数的极大似然估计同样可以用式(3.11)和式(3.12)两式表示。

由于$\ln x$是x的单调增函数,使

$$\ln L(\theta;X_1,\cdots,X_n)=\sup_{\theta\in\Theta}\ln L(\theta;X_1,\cdots,X_n) \tag{3.14}$$

成立的$\hat{\theta}$也使式(3.12)成立,所以有时只要从式(3.14)中求$\hat{\theta}$即可。

简单地说,极大似然估计的定义就是:

满足条件$L(\hat{\theta})=\max_{\theta}\{L(\theta)\}$的$\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$叫做参数$\theta$的极大似然估计值。其中$L(\cdot)$表示样本的似然函数。

求极大似然函数估计值的一般步骤如下:

(1) 写出似然函数。

(2) 对似然函数取对数,并整理。

(3) 求导数。

(4) 解似然方程。

极大似然估计只是一种概率论在统计学中的应用,它是参数估计的方法之一。说的是已知某个随机样本满足某种概率分布,但是其中具体的参数不清楚,参数估计就是通过若干次试验,观察其结果,利用结果推出参数的大概值。极大似然估计是建立在这样的思想上:已知某个参数能使这个样本出现的概率最大,我们当然不会再去选择其他小概率的样本,所以干脆就把这个参数作为估计的真实值。

当然极大似然估计只是一种粗略的数学期望，要知道它的误差大小还要做区间估计。

3. 基于极大似然参数估计的贝叶斯网络参数学习

具体到贝叶斯网络，似然函数可以表示成

$$L(\boldsymbol{\theta}:D)=\prod_{m=1}^{M}P(X_1[m],X_2[m],\cdots,X_n[m]\mid\boldsymbol{\theta})=\prod_{m=1}^{M}\prod_{i=1}^{n}P(X_i[m]\mid \mathrm{Pa}_{X_i},\boldsymbol{\theta}_i)$$
$$=\prod_{i=1}^{n}\prod_{m=1}^{M}P(X_i[m]\mid \mathrm{Pa}_{X_i},\boldsymbol{\theta}_i)=\prod_{i=1}^{n}\prod_{j=1}^{q}L(\boldsymbol{\theta}_{ij}:D) \tag{3.15}$$

要求 $L(\boldsymbol{\theta}:D)$的极大值，即是要求每个因子 $L(\boldsymbol{\theta}_{ij}:D)$的极大值；同时根据贝叶斯网络的性质，有式(3.16)成立，即

$$L(\boldsymbol{\theta}_{ij}:D)=P(D\mid\boldsymbol{\theta}_{ij})=\prod_{k=1}^{r_i}\theta_{ijk}{}^{N_{ijk}} \tag{3.16}$$

式中，r_i 为 X_i 的取值数目；N_{ijk}为第 i 个结点，取值为第 k 个值时，在父结点取值组合取第 j 个值时数据的数量，该值被称为充分统计因子，$\{N_{ijk}\mid i=1,\cdots,n;j=1,\cdots,q;k=1,\cdots,r_k\}$是充分统计量(Sufficient Statistics)。

按照最大似然估计值的求法，令

$$\frac{\partial(P(D\mid\boldsymbol{\theta}_{ij}))}{\partial\,\boldsymbol{\theta}_{ij}}=0$$

即

$$\frac{\partial\left(\prod_{k=1}^{r_i}\theta_{ijk}{}^{N_{ijk}}\right)}{\partial\,\boldsymbol{\theta}_{ij}}=0 \tag{3.17}$$

通过计算，最终可以得到参数的计算公式为

$$\hat{\theta}_{ijk}=\frac{N_{ijk}}{N_{ij}}$$

其中

$$N_{ij}=\sum_{k=1}^{r_i}N_{ijk} \tag{3.18}$$

3.3.2 贝叶斯参数估计

1. 概念

贝叶斯参数估计(Bayesian Parameter Estimation)除了利用总体信息、样本信息之外，还利用试验之前已经获得的信息——先验信息，对总体参数进行估计。根据样本分布和未知参数的先验分布，用概率论中求条件概率分布的方法，求出的在样本已知情况下，未知参数的条件分布。因为这个分布是在抽样以后才得到的，故称为后验分布。贝叶斯推断方法的关键是任何推断都必须且只需根据后验分布，而不能再涉及样本分布。下面先介绍一下先验分布、后验分布和贝叶斯公式。

1) 先验分布

设 Θ 是参数空间，$\theta\in\Theta$ 是未知参数，称先验信息的数量化为参数空间 Θ 上确定先验分

布，记为 $\pi(\theta)$。

可用于确定先验分布的方法有以下几种：

(1) 主观概率。按贝叶斯学派的说法，这是一种通过“自我反省”去确定先验分布的方法。就是说，对参数 θ 取某某值的可能性多大，通过思考，觉得该如何，而确定一个值。

主观先验分布反映了个人以往对 θ 的了解，包括经验知识和理论知识，其中有部分可能是通过他人获取的，也可能是他人对 θ 的了解。

对过去的经验和知识，必须经过组织和整理。这样提出的先验分布，在主观上是正确的，但不能保证符合客观标准。

例如，若对事件 A 与 $\overline{A}$，根据经验可知 A 发生的可能性是 $\overline{A}$ 发生的可能性的 k 倍，即

$$P(A) = kP(\overline{A}) \tag{3.19}$$

则有事件 A 的主观概率

$$P(A) = \frac{k}{k+1} \tag{3.20}$$

(2) 无信息先验分布。即参数 θ 可以取某个区域内的任意一值。

若 Θ 是区间 $[a,b]$，则可认为 θ 服从 $[a,b]$ 上的均匀分布。

这一原则称为贝叶斯假定。以产品的废品率 p 为例，当对 p 一无所知时，只好先验地认为，p 以同等机会取(0,1)内各种值，因而以(0,1)内均匀分布 $U(0,1)$ 作为 p 的先验分布。需要注意的是，这一原则会出现矛盾。如果对 p 无知，对 p^3 也同样无知。按同等无知原则，可以取 $U(0,1)$ 作为 p^3 的分布，但这时 p 的分布就不是 $U(0,1)$ 了。

(3) 共轭分布方法。H. Raiffa、R. Schlaifer 提出先验分布应取共轭分布才合适，设样本 X 的分布族为 $\{f(x|\theta):\theta\in\Theta\}$，若先验分布 $\pi(\theta)$ 与后验分布 $\pi(\theta|x)$ 属于同一分布类型，则先验分布 $\pi(\theta)$ 称为 $\pi(\theta|x)$ 的共轭分布，Dirichlet 分布为常用的共轭分布之一。

(4) Jeffreys 原则。Jeffreys 提出的选取先验分布的原则是一种不变原理，较好地解决了贝叶斯假设中的一个矛盾，即若对参数 θ 选用均匀分布，则其函数 $g(\theta)$ 往往不是均匀分布。

Jeffreys 原则：设按照原则决定 θ 的先验分布为 $h(\theta)$，若以 $g(\theta)$ 作为参数，按同一原则决定的 $\eta=g(\theta)$ 的先验分布是 $h_g(\eta)$，则应用关系式

$$h(\theta) = h_g[g(\theta)] \mid g'(\theta) \mid \tag{3.21}$$

若选取的 $h(\theta)$ 符合式(3.21)，则用 θ 或 θ 的函数 $g(\theta)$ 的先验分布总是一致的。

这个方法的困难之处在于如何找到满足式(3.21)的 $h(\theta)$，Jeffreys 利用 Fisher 信息量的不变性，找到了符合要求的 $h(\theta)$。

(5) 最大熵原则。熵是信息论的一个基本概念，是随机变量不确定性的度量。不确定性越大，则熵越大，在“无信息”的情况下，应取熵最大的分布为先验分布。

最大熵原则：无信息先验分布应取参数 θ 的变化范围内使熵最大的分布。

2) 后验分布与贝叶斯公式

在 $X=x$ 的条件下，θ 的条件分布称为 θ 的后验分布，后验分布由后验概率密度函数来描述。

设参数 θ 的先验信息已知的条件下，总体 X 的条件分布密度函数是 $f(x|\theta)$，X 的样本 $(X_1,X_2,\cdots,X_n)$ 的条件分布密度函数为

$$f(x_1,x_2,\cdots,x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{3.22}$$

设θ的先验分布是$\pi(\theta)$，则$(X_1,X_2,\cdots,X_n,\theta)$的联合分布密度函数为

$$\varphi(x_1,x_2,\cdots,x_n,\theta)=\pi(\theta)f(x_1,x_2,\cdots,x_n\mid\theta) \tag{3.23}$$

那么$(X_1,X_2,\cdots,X_n,\theta)$关于$(X_1,X_2,\cdots,X_n)$的边缘分布密度函数为

$$g(x_1,x_2,\cdots,x_n)=\int_{\Theta}\varphi(x_1,x_2,\cdots,x_n,\theta)\mathrm{d}\theta \tag{3.24}$$

当样本观测值已知后，$(X_1,X_2,\cdots,X_n,\theta)$的联合分布密度函数可以分解为

$$\varphi(x_1,x_2,\cdots,x_n,\theta)=g(x_1,x_2,\cdots,x_n)\pi(\theta\mid x_1,x_2,\cdots,x_n) \tag{3.25}$$

称

$$\pi(\theta\mid x_1,x_2,\cdots,x_n)=\frac{\varphi(x_1,x_2,\cdots,x_n,\theta)}{g(x_1,x_2,\cdots,x_n)}=\frac{\pi(\theta)f(x_1,x_2,\cdots,x_n\mid\theta)}{\int_{\Theta}\pi(\theta)f(x_1,x_2,\cdots,x_n\mid\theta)\mathrm{d}\theta} \tag{3.26}$$

为分布密度函数形式的贝叶斯公式。

$\pi(\theta|x_1,x_2,\cdots,x_n)$是样本$(X_1,X_2,\cdots,X_n)$的取值给定时关于$\theta$的条件分布密度函数，是获得样本观测值后对先验信息的补充或改进。称$\pi(\theta|x_1,x_2,\cdots,x_n)$为$\theta$的后验分布。

2. 贝叶斯参数估计原理

贝叶斯学派认为：后验分布族是统计推断的出发点。这里样本观测值是确定的，而θ是随机的。而经典统计中，其出发点是样本分布族，其中θ是未知参数，而样本观测值x只是无限次可能试验结果的一个具体实现。总体来说，样本X是随机的。

无偏性与样本分布族有关，因而“无偏性”不符合贝叶斯统计推断原则。

总之，贝叶斯统计推断的任务是根据已知的样本观测值x对未知的随机变量根据后验分布作出推断，这里x是具体的值，没有必要将其放在“无限多可能值之一”中考察。

参数θ的贝叶斯估计有以下几种方法：

(1) 用后验期望(后验均值)$\hat{\theta}$估计θ

$(X_1,X_2,\cdots,X_n)$的取值已知，设θ的后验密度为$\pi(\theta|\vec{X})$，$\vec{X}=(x_1,x_2,\cdots,x_n)$，$\theta$的后验期望为

$$\hat{\theta}=E(\theta\mid\vec{X})=\int\theta\cdot\pi(\theta\mid\vec{X})\mathrm{d}\theta \tag{3.27}$$

用$\hat{\theta}=E(\theta|\vec{X})$去估计$\theta$是一个很自然的想法。

(2) 最小均方误差估计

设$\hat{\theta}$是θ的估计量，$E(\hat{\theta}-\theta)^2$称为均方误差，最小均方误差就是求$\hat{\theta}$，使

$$\hat{\theta}=\arg\min_{\theta}E(\hat{\theta}-\theta)^2 \tag{3.28}$$

在贝叶斯统计中，$\hat{\theta}$与θ都是随机变量，$E(\hat{\theta}-\theta)^2$的计算与传统的不同，有

$$E(\hat{\theta}-\theta)^2=\iint(\hat{\theta}-\theta)^2\cdot\pi(\theta\mid\vec{X})\mathrm{d}x\mathrm{d}\theta \tag{3.29}$$

(3) 最大后验估计

用一组样本集$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$估计未知参数$\theta$，未知参数$\theta$视为随机变量，先验分布为$\pi(\theta)$，而在已知样本集$\boldsymbol{X}$出现的条件下的后验概率为$\pi(\theta|\boldsymbol{X})$，则最大后验概率估计(Maximum a posteriori，MAP)为

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} \pi(\theta \mid \boldsymbol{X}) = \arg\max_{\theta} \frac{\pi(\boldsymbol{X} \mid \theta)\pi(\theta)}{\pi(\boldsymbol{X})}$$
$$= \arg\max_{\theta} \pi(\boldsymbol{X} \mid \theta)\pi(\theta) \tag{3.30}$$

对于给定总体中的某一个待估计参数，不同的估计方法可以得到该参数不同的估计量。因此，需要对估计量的优劣作出评价。评价估计量优劣的标准主要有无偏性、有效性、相合性（亦称一致性）。需特别说明的是，除了采用无偏、有效和相合估计作为评价准则外，还可以利用估计误差的变化范围作为参数估计的测度，这种测度叫做损失函数（或代价函数）。常用于标量估计的贝叶斯估计器的损失函数有绝对型、二次型和均匀型。

一般贝叶斯估计，设 $\varepsilon=\theta-\hat{\theta}$ 表示估计误差，令 $C(\varepsilon)$ 为损失函数，定义 $R=E[C(\varepsilon)]$ 为贝叶斯风险函数。令贝叶斯风险函数最小，计算得到各种贝叶斯估计。

绝对型损失函数为

$$C(\varepsilon) = |\hat{\theta} - \theta| \tag{3.31}$$

二次型损失函数为

$$C(\varepsilon) = \varepsilon^2 \tag{3.32}$$

均匀型损失函数为

$$C(\varepsilon) = \begin{cases} 0 & |\varepsilon| < \delta \\ 1 & |\varepsilon| > \delta \end{cases} \tag{3.33}$$

其风险函数的解是 MAP 估计。

3\. 基于最大后验分布估计的贝叶斯网络参数学习

由于最大似然估计法没有利用先验知识，故收敛速度较慢，而基于贝叶斯统计的方法则恰恰利用了这部分知识。

在以前的分析过程中，对完整数据集的参数，有各个参数向量$\boldsymbol{\theta}_{ij}$相互独立的假设；同时可以证明，参数$\boldsymbol{\theta}_{ij}$的后验分布也具有相同的独立性，所以进行参数估计时，可以分别计算基于各个参数的后验分布 $P(\boldsymbol{\theta}_{ij}|D)$，并在此基础上进行参数估计。

在此，把先验分布选为 Dirichlet 分布，即

$$\mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha'_{ij1}, \alpha'_{ij2}, \cdots, \alpha'_{ijr_i}) = \frac{\Gamma(\alpha'_{ij})}{\prod_{k=1}^{r_i} \Gamma(\alpha'_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}\, \alpha'_{ijk} - 1 \tag{3.34}$$

该公式中的变量可以参考 BDe 评分函数部分，则$\boldsymbol{\theta}_{ij}$的后验分布为：

$$P(\boldsymbol{\theta}_{ij} \mid D) = \mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha'_{ij1} + \alpha_{ij1}, \alpha'_{ij2} + \alpha_{ij2}, \cdots, \alpha'_{ijr_i} + \alpha_{ijr_i})$$
$$= \frac{\Gamma(\alpha'_{ij} + \alpha_{ij})}{\prod_{k=1}^{r_i} \Gamma(\alpha'_{ijk} + \alpha_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha'_{ijk} + \alpha_{ijk} - 1} \tag{3.35}$$

采用求数学期望的方法实现对参数的估计，即

$$\hat{\theta}_{ijk} = E_{P(\boldsymbol{\theta}_{ij} \mid D)}(\theta_{ijk}) = \frac{\alpha'_{ijk} + \alpha_{ijk}}{\sum_{k=1}^{r_i} (\alpha'_{ijk} + \alpha_{ijk})} \tag{3.36}$$

例如，对射击队员的命中率进行估计，第一个队员射击数为10次，命中6次；第二个队员射击100次，命中60次。按照最大似然方法估计两次的命中率均为$\frac{6}{10}=\frac{60}{100}=60\%$，两个射击队员的水平相当。假设先验信息为无信息先验分布，即对两名射击队员的水平预先不知，根据贝叶斯假设，使用均匀分布作为其分布，即参数在它的变化范围内，取到各个值的机会是相同的。根据上述的讨论，取$\alpha'_{ijk}=1$，第一个射击队员的命中率为$\frac{1+6}{2+10}=0.583$，第二个射击队员的命中率为$\frac{1+60}{2+100}=0.598$。显然，人们对第二个射击队员的命中率更加充满信心，使用Dirichlet为先验知识进行评估也更加合理。

另外，还可以计算D中第$m+1$个实例出现的概率，即

$$
\begin{aligned}
P(X^{N+1}\mid D,S) &= \prod_{i=1}^{n}\int_{\theta_{ij1}}\cdots\int_{\theta_{ijr_i}}\theta_{ijk}P(\theta_{ij}\mid D)\mathrm{d}\theta_{ij} \\
&= \prod_{i=1}^{n}\int_{\theta_{ij1}}\cdots\int_{\theta_{ijr_i}}\theta_{ijk}\mathrm{Dir}(\theta_{ij}\mid \alpha'_{ij1}+\alpha_{ij1},\alpha'_{ij2}+\alpha_{ij2},\cdots,\alpha'_{ijr_i}+\alpha_{ijr_i})\mathrm{d}\theta_{ij} \\
&= \prod_{i=1}^{n}\frac{\alpha'_{ijk}+\alpha_{ijk}}{\alpha'_{ij}+\alpha_{ij}}
\end{aligned}
\tag{3.37}
$$

其中$\alpha'_{ij}=\sum_{k=1}^{r_i}\alpha'_{ijk}$、$\alpha_{ij}=\sum_{k=1}^{r_i}\alpha_{ijk}$，$\alpha'_{ijk}$代表专家知识。在实际计算中，可以采用等价抽样法来估计；在缺乏专家知识时，也可以采用$\alpha'_{ijk}=1$来估计。

第4章
不完整数据集下的有向概率图模型的学习

4.1 概述

由于各种原因，现实数据库中经常会存在不同程度的数据丢失现象，具有缺失数据的有向概率图模型学习一直是近年来的研究热点和难点。数据的缺失主要会产生两方面的问题：一方面，评分函数不再具有可分解形式，不能进行局部搜索；另一方面，一些充分统计因子不存在，无法直接进行结构评分。围绕这两个问题，相继出现了一些解决方法，目前，对于具有缺失数据的贝叶斯网络的参数学习，常用的方法有 Gibbs 抽样方法和 EM (Expectation-Maximization)方法。对于具有缺失数据的贝叶斯网络的结构学习，主要有结构 EM(Structural EM，SEM)方法和评分结构方法。

本章基于贝叶斯网络，对参数学习的 EM 算法与 Gibbs 抽样方法、结构学习的 SEM 算法和评分—搜索结构方法作简要介绍。

4.2 参数估计

数据集不完整，是指数据在某些结点处有缺失现象。此时，已不能再利用参数之间的独立性进行计算，即不能将似然函数或后验概率分解成关于每个参数的因子，前边介绍的参数学习算法将不再适用。对于不完整数据集，常采用的方法有 EM 算法、Gibbs 抽样方法等。

4.2.1 EM 算法

在人工智能、数理统计、机器学习和模式识别中有许多的应用都要进行模型的参数估计，也就是要进行极大似然估计(MLE)或极大后验似然估计(MAP)。当模型中的变量均为可以直接观察的变量时，极大似然或极大后验似然是显然的。但是当某些变量隐藏时，进行极大似然估计就比较复杂。也就是说，在完整数据情况下，贝叶斯网络的参数最大似然估计可以用式(3.18)来计算。但是当数据有缺值时，$P(D_i|\theta)$不能按式(3.16)分解，因而没有计算最大似然估计的公式。

实际上，在存在潜在变量或数据不完全的情况下，人们提出了很多对模型参数进行估计的方法，一种非常流行的极大似然估计方法是 EM 算法。它不是直接对复杂的后验分布进行极大化或模拟，而是在观察数据的基础上添加一些“潜在数据”，从而简化计算并完成一系列简单的极大化或模拟。它之所以被称为 EM 算法是因为算法的每一次迭代由一个期望

步(E-step)和极大步(M-step)构成。EM 算法的特点是简单和稳定，特别是每一次迭代能保证观察数据对数的后验似然是单调不减的。

EM 算法是一种一般的从“不完全数据”中求解模型参数的极大似然估计的方法。“不完全数据”一般分为两种情况：一种是由于观察过程本身的限制或者错误，造成观察数据成为错漏的不完全数据；另一种是参数的似然函数直接优化十分困难，而引入额外的参数(隐含的或丢失的)后就比较容易优化，是定义原始观察数据加上额外数据组成“完全数据”，原始观察数据自然就成为“不完全数据”。

EM 算法为了求得贝叶斯网络 $B=(G,\theta)$ 的参数 θ 的 MLE 的近似值，其基本方法是：从 θ 的某个初始值 θ^0 开始迭代，初始值 θ^0 往往是随机产生的。设已经进行了 t 次迭代，得到估计 θ^t，第 $t+1$ 次迭代由以下两个步骤组成：

(1) E-step。利用网络结构和当前参数 θ^t 对缺失数据计算它的期望值，使数据修补完整。

(2) M-step。基于修补后的完整数据，计算出新的最大可能的参数分布 θ^{t+1}，用 θ^{t+1} 替换原有的 θ^t。

重复上述过程，直到所估计的参数达到局部最优或达到指定的迭代次数。

下面介绍这两个步骤的具体实现过程。E-step 步骤是进行数据的修补，在这个过程中，考虑某一缺值样本 $D[m]$，设 X_m 为 $D[m]$ 中所有缺值变量的集合。对 X_m 的一个取值 x_m，将 $X_m=x_m$ 加入 $D[m]$ 就得到一完整样本，由于 X_m 有多个可能取值，所以 $D[m]$ 有多种修补方式，EM 算法考虑所有可能的修补结果，根据当前估计 θ^t 与网络结构，计算 $P(X_m=x_m \mid D[m],\theta^t)$，所得的结果为样本 $<D[m],X_m=x_m>$ 的权重，记为 $Q_m(<X_m=x_m>)$，这样带有权重的样本又称为碎权样本(Fractional Sample)。在数据修补过程中，每个缺值样本都被一系列完整的碎权样本所代替。所以，在修补后的数据中所有样本都是完整的，而且每个样本都有一个权重，原本完整的样本权重为 1。修补完样本后，M-step 步骤基于修补的完整数据，计算出新的最大可能的参数分布 θ^{t+1}。与式(3.18)类似，EM 算法用式(4.1)计算基于补后数据的最大似然估计 θ^{t+1}，即

$$\theta_{ijk}^{t+1}=\begin{cases}\dfrac{\overline{M}_{ijk}}{\sum\limits_{k=1}^{r_i}\overline{M}_{ijk}}, & 若 \sum\limits_{k=1}^{r_i}\overline{M}_{ijk}>0\\[2ex] \dfrac{1}{r_i}, & 若否\end{cases} \tag{4.1}$$

式中，$\overline{M}_{ijk}$ 为补后数据中所有满足 $X_i=k$ 和其父结点集合 $\mathrm{Pa}(X_i)=j$ 的样本的权重之和。

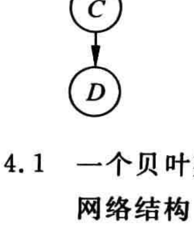


图 4.1 一个贝叶斯网络结构

例如，有如图 4.1 所示的贝叶斯网络结构，其所有结点变量为布尔变量，现在有不完整数据 $D[1]=(a^1,?,?,d^0)$ 和 $D[2]=(?,b^1,?,d^1)$。假如，当前的 θ 是

$$\theta_{a^1}=0.4 \qquad \theta_{b^1}=0.8$$
$$\theta_{d^1|c^0}=0.2 \qquad \theta_{d^1|c^1}=0.7$$
$$\theta_{c^1|a^0,b^0}=0.7 \qquad \theta_{c^1|a^1,b^0}=0.8$$
$$\theta_{c^1|a^0,b^1}=0.1 \qquad \theta_{c^1|a^1,b^1}=0.3$$

则首先 E-step 对数据进行修补。对于 $D[1]=(a^1,?,?,d^0)$，其碎权样本的权重计算为

$$Q_1(<b^1,c^1>)=P(b^1,c^1\mid a^1,d^0)=\frac{P(b^1,c^1,a^1,d^0)}{P(a^1,d^0)}$$

$$=\frac{P(a^1)P(b^1)P(c^1\mid a^1,b^1)P(d^0\mid c^1)}{P(a^1,d^0)}$$

$$=0.4\cdot 0.8\cdot 0.3\cdot 0.3/0.24=0.12$$

$$Q_1(<b^1,c^0>)=P(b^1,c^0\mid a^1,d^0)=\frac{P(b^1,c^0,a^1,d^0)}{P(a^1,d^0)}$$

$$=\frac{P(a^1)P(b^1)P(c^0\mid a^1,b^1)P(d^0\mid c^0)}{P(a^1,d^0)}$$

$$=0.4\cdot 0.8\cdot 0.7\cdot 0.8/0.24=0.7467$$

$$Q_1(<b^0,c^1>)=P(b^0,c^1\mid a^1,d^0)=\frac{P(b^0,c^1,a^1,d^0)}{P(a^1,d^0)}$$

$$=\frac{P(a^1)P(b^0)P(c^1\mid a^1,b^0)P(d^0\mid c^1)}{P(a^1,d^0)}$$

$$=0.4\cdot 0.2\cdot 0.8\cdot 0.3/0.24=0.08$$

$$Q_1(<b^0,c^0>)=P(b^0,c^0\mid a^1,d^0)=\frac{P(b^0,c^0,a^1,d^0)}{P(a^1,d^0)}$$

$$=\frac{P(a^1)P(b^0)P(c^0\mid a^1,b^0)P(d^0\mid c^0)}{P(a^1,d^0)}$$

$$=0.4\cdot 0.2\cdot 0.2\cdot 0.8/0.24=0.0533$$

对于 $D[2]=(?,b^1,?,d^1)$，其碎权样本的权重计算为

$$Q_2(<a^1,c^1>)=P(a^1,c^1\mid b^1,d^1)=\frac{P(a^1,c^1,b^1,d^1)}{P(b^1,d^1)}$$

$$=\frac{P(a^1)P(b^1)P(c^1\mid a^1,b^1)P(d^1\mid c^1)}{P(b^1,d^1)}$$

$$=0.4\cdot 0.8\cdot 0.3\cdot 0.7/0.232=0.2897$$

$$Q_2(<a^1,c^0>)=P(a^1,c^0\mid b^1,d^1)=\frac{P(a^1,c^0,b^1,d^1)}{P(b^1,d^1)}$$

$$=\frac{P(a^1)P(b^1)P(c^0\mid a^1,b^1)P(d^1\mid c^0)}{P(b^1,d^1)}$$

$$=0.4\cdot 0.8\cdot 0.7\cdot 0.2/0.232=0.1931$$

$$Q_2(<a^0,c^1>)=P(a^0,c^1\mid b^1,d^1)=\frac{P(a^0,c^1,b^1,d^1)}{P(b^1,d^1)}$$

$$=\frac{P(a^0)P(b^1)P(c^1\mid a^0,b^1)P(d^1\mid c^1)}{P(b^1,d^1)}$$

$$=0.6\cdot 0.8\cdot 0.1\cdot 0.7/0.232=0.1448$$

$$Q_2(<a^0,c^0>)=P(a^0,c^0\mid b^1,d^1)=\frac{P(a^0,c^0,b^1,d^1)}{P(b^1,d^1)}$$

$$=\frac{P(a^0)P(b^1)P(c^0\mid a^0,b^1)P(d^1\mid c^0)}{P(b^1,d^1)}$$

$$=0.6\cdot 0.8\cdot 0.9\cdot 0.2/0.232=0.3724$$

计算出以上碎权样本权重以后，原先仅有两个不完整样本的数据集被修补为 8 个完整样本的数据集，如表 4.1 所示。

表 4.1 E-step 修补后的完整数据集

	X_1	X_2	X_3	X_4	Q
$D'[1]$	a^1	b^1	c^1	d^0	0.12
$D'[2]$	a^1	b^1	c^0	d^0	0.7467
$D'[3]$	a^1	b^0	c^1	d^0	0.08
$D'[4]$	a^1	b^0	c^0	d^0	0.0533
$D'[5]$	a^1	b^1	c^1	d^1	0.2897
$D'[6]$	a^1	b^1	c^0	d^1	0.1931
$D'[7]$	a^0	b^1	c^1	d^1	0.1448
$D'[8]$	a^0	b^1	c^0	d^1	0.3724

修补完数据后，M-step 要计算出新的最大可能的参数分布 θ'，用 θ' 替换原有的 θ。假设要计算新的 $\theta'_{d^1|c^0}$，根据式(4.1)有

$$\begin{aligned}\theta'_{d^1|c^0} &= \frac{\overline{M}_{d^1|c^0}}{\sum\limits_{d^0,d^1}\overline{M}_{c^0}} \\ &= \frac{Q_2(<a^1,c^0>)+Q_2(<a^0,c^0>)}{Q_1(<b^1,c^0>)+Q_1(<b^0,c^0>)+Q_2(<a^1,c^0>)+Q_2(<a^0,c^0>)} \\ &= \frac{0.1931+0.3724}{0.7467+0.08+0.1931+0.3724} \\ &= 0.4062\end{aligned}$$

可以依次类推，计算出其他参数新的估计值，并且用新的参数估计值代替旧的参数，不断重复 E-step 和 M-step，直到收敛。

下面给出了贝叶斯网络参数的 EM 学习算法的伪代码：

Algorithm4.1 Expectation-maximization algorithm for BN

//计算 E-step 中的期望充分统计量

Compute-ESS (

G, //贝叶斯网络结构

θ, //贝叶斯网络参数

D, //不完整数据集

)

1 //初始化数据结构

2 For each $i=1,\cdots,n$

3 For each $x_i,u_i\in \mathrm{Val}(X_i,\mathrm{Pa}_{X_i}^G)$

4 $\overline{M}_{x_i|u_i}\leftarrow 0$

5 For each $m=1,\cdots,M$

6 Run inference on (G,θ) using evidence $D[m]$ //使用数据样本 $D[m]$ 在网络结构和当前的参数上进行推理运算

7 For each $i=1,\cdots,n$

8 For each $x_i,u_i\in \mathrm{Val}(X_i,\mathrm{Pa}_{X_i}^G)$

9 $\overline{M}_{x_i|u_i}\leftarrow \overline{M}_{x_i|u_i}+p(x_i,u_i\mid D[m])$

```
10 Return {M̄_{x_i|u_i} : ∀ i=1, … ,n, ∀ x_i, u_i ∈ Val(X_i, Pa^G_{X_i})}
Expectation-Maximization(
G,  //贝叶斯网络结构
θ^0,  //贝叶斯网络初始参数
D,  //不完整数据集
)
1 For each t=0,1, … , until convergence
2   //E-step
3 {M̄^t_{x_i|u_i}}←Compute-ESS(G, θ^t, D)
4   //M-step
5 For each i=1, … ,n
6     For each x_i, u_i ∈ Val (X_i, Pa^G_{X_i})
7 θ^{t+1}_{x_i|u_i} ← M̄^t_{x_i|u_i} / M̄^t_{u_i}
8 Return θ^t
```

4.2.2 Gibbs 抽样方法

Gibbs 抽样方法是由 Heckerman 提出来用于不完备数据参数学习的方法之一。在贝叶斯推理中，Gibbs 是最为流行的马尔可夫蒙特卡罗(MCMC)方法之一。Gibbs 把含有不完备数据集 D 的每一个缺项当作待估计参数，通过对未知参数后验分布的一系列随机抽样过程，计算参数的后验均值的经验估计。

Gibbs 抽样法的基本思想是利用已有数据推断丢失的数据，使具有丢失数据的数据库成为完整的数据库。其近似参数后验的概率分布的基本步骤如下：

① 随机初始化每个数据样本中的缺值变量，得到一完整的数据集 D_c。

② 选择某个在原数据集中缺失的变量值 x_{il}(变量 x_i 在第 l 个样本中的取值)，去掉其初始化的值，并利用下面的公式计算概率分布：

$$P(x'_{il} \mid D_c/x_{il}) = \frac{P(x'_{il}, D_c/x_{il})}{\sum_{x''_{il}} P(x''_{il}, D_c/x_{il})} \tag{4.2}$$

式中，D_c/x_{il} 为在 D_c 中去掉 x_{il} 值的数据集；x'_{il} 为某一种状态；x''_{il} 为所有可能的状态；分母是对 x_{il} 的所有可能情况的求和，分子和分母利用完备数据集的似然计算公式进行计算。

③ 根据概率分布对丢失数据进行修正，对数据集中每一个缺失值，重复步骤②，从而得到新的完备数据集 D'_c。

④ 利用完备数据集 D'_c，计算参数后验的概率分布 $P(\theta|D'_c, S)$。

⑤ 重复前 4 个步骤。

⑥ 将每次迭代的参数估计 $P(\theta|D'_c, S)$ 的均值作为最终的参数估计。

4.3 结构学习

4.3.1 结构 EM 算法

结构 EM(Structural EM, SEM)算法是贝叶斯网络缺值数据下的结构学习的主要算

法。Friedman对贝叶斯网络的结构学习提出了SEM算法，其中包含与EM相同的E-step，SEM的M-step是根据E-step修补后的数据和当前的网络结构，评价任意一个候选网络结构，其网络结构的评价函数可以是包括BIC评分在内的任意一个评分函数，那么最终的结构应该是分值最高的。

它的基本思想是：从某初始模型结构G^0和参数θ^0出发开始迭代，在进行了t次迭代得到了(G^t,θ^t)后，第$t+1$次迭代由以下两个步骤组成：

(1) 基于(G^t,θ^t)对数据进行修补，使之完整。

(2) 基于修补后的完整数据D^t对模型及参数进行一步优化，得到(G^{t+1},θ^{t+1})。

在SEM中，基于数据D^t进行一次优化未必得到基于D^t的最优模型结构，因为在完整数据下，模型结构优化不是一步就完成的。所以，SEM不是每次迭代都同时优化模型结构和参数，而是先固定模型结构，即规定$G^{t+1}=G^t$，进行数次参数优化后，再进行一次结构加参数优化，如此交替进行。

SEM算法的伪代码如下：

Algorithm4.2 The structural EM algorithm for BN structure learning

SEM(
G^0, //初始网络结构
$\theta^{0,0}$, //初始参数值
D, //缺值数据集
R //两次结构优化之间的参数优化次数
)
1 For $t=0,\cdots$, until convergence
2 For r=0 to R−1
3 $\theta^{t,r+1} \leftarrow$ Expectation-Maximization(G^t, $\theta^{t,r}$, D)
4 $\theta^t=\theta^{t,R}$
5 $L \leftarrow$ 所有对G^t做一次加边、减边或转边而得到的候选模型结构
6 $G^{t+1} \leftarrow \arg\max_{G \in L} \mathrm{Score}(G|D^t,\theta^t)$
7 $\theta^{t+1,0} \leftarrow$ 估计G^{t+1}的参数值
8 If(Score(G^{t+1}, $\theta^{t+1,0}$ | D) $\leqslant$ Score(G^t, θ^t | D))
9 Return(G^t, θ^t)
10 Return(G^{t+1}, θ^{t+1})

SEM算法先进行R次参数优化(2,3两行)，接着同时优化模型结构及参数(6、7两行)。如果模型加参数优化得到评分更高的模型，SEM就重复前面的运算，否则就停止并返回找到得评分最高的贝叶斯网(8、9两行)。

对于DBNs的结构学习，Friedman对SEM算法进行了扩展。

设$L=\{(X_i[0],Y[0]),\cdots,(X_i[T],Y[T])\}$为少量完整数据序列集，$U=\{(X_j[0]),\cdots,(X_j[T])\}$为不完整数据序列集，其DBNs-SEM算法描述如下：

Algorithm4.3 The structural EM algorithm for DBN learning

DBNs-SEM(
L, //少量完整数据集

U //大量完整数据集

)

1 使用 L 学习一个初始动态贝叶斯网络$(B_0^0, B_{\rightarrow}^0)$

2 For n=0,1,… 直到满足停止条件

3 使用 L 和 U,利用 EM 算法对 DBNs 进行参数学习,获得参数提高$(B_0^n, B_{\rightarrow}^n)$和修补后数据 U'

4 基于 L 和 U',对$(B_0^n, B_{\rightarrow}^n)$的每个候选网络结构计算评分函数

5 选择评分最高的网络结构作为$(B_0^{n+1}, B_{\rightarrow}^{n+1})$

6 如果 $((B_0^{n+1}, B_{\rightarrow}^{n+1}) \leqslant (B_0^n, B_{\rightarrow}^n))$,那么返回$(B_0^n, B_{\rightarrow}^{n1})$

7 返回 $(B_0^{n+1}, B_{\rightarrow}^{n+1})$

4.3.2 打分-搜索方法

在 3.2 节中曾经介绍了在完整数据集下应用 BIC 评分函数和 BDe 评分函数来学习网络结构。那么当数据缺失情况下,怎样运用这些评分函数进行结构学习呢?可以这样来看,想要的评分同在完整数据下的评分是一样的,有

$$\text{score}_B(G:D) = \log P(D \mid G) + \log P(G) \tag{4.3}$$

这里 $P(D|G)$是数据的边界似然,即

$$P(D \mid G) = \int_{\theta_G} P(D \mid \theta_G, G) P(\theta_G \mid G) \mathrm{d}\theta_G \tag{4.4}$$

在完整数据集下,$P(D|G)$的积分计算内部是可分解的乘法因式,可以被简化。但是,在不完整数据集下,这个后验概率是不能被分解的。

然而,可以把 $P(D|G)$的计算看作是一个推理问题。对于大多数的不完整数据学习问题,推理是很难处理的,通常可以采取 Laplace 近似、Asymptotic 近似、Cheeseman-Stutz 近似、候选方法等不同策略来解决[298]。

下面简单介绍一下候选方法的基本思想。

候选方法利用参数的候选集来估计 $P(D|G)$。对于任意一组参数的集合,利用概率的链规则,可以写出两种不同的 $P(D,\theta|G)$表达式,即

$$\begin{aligned} P(D,\theta \mid G) &= P(D \mid \theta, G) P(\theta \mid G) \\ P(D,\theta \mid G) &= P(\theta \mid D, G) P(D \mid G) \end{aligned} \tag{4.5}$$

两式右边相等,可以推出

$$P(D \mid G) = \frac{P(D \mid \theta, G) P(\theta \mid G)}{P(\theta \mid D, G)} \tag{4.6}$$

式中,分子上的 $P(D|\theta,G)$为给定 θ 后观测数据的似然,其可以用精确推理或近似推理来估计;$P(\theta|G)$为参数的先验,通常已给定;分母上的 $P(\theta|D,G)$是参数的后验,可以使用 MCMC 抽样参数候选集的方法来估计。当然估计的准确与否,决定于 MCMC 抽样器的设计,要求估计值越准确,那么相应的 MCMC 抽样计算量也增加,其完全可以达到精确估计 $P(\theta|D,G)$,从而得到精确估计的 $P(D|G)$。

确定计算网络结构的评分策略以后,可以用搜索算法对候选网络进行打分,找出评分最高的网络作为最终的网络结构。但是对于像爬山法一类的搜索算法,并不能直接应用,因为这类方法依赖于 $P(D|G)$的可分解性。

第 5 章
无向概率图模型学习

5.1 概述

在第 3 章和第 4 章中，介绍的都是有向概率图模型，或者说是贝叶斯网络的学习方法，这些模型能够从结构和参数上自然表达现实世界的一些典型问题。然而在本章中，将关注另一种概率图模型的学习，即无向概率图模型的学习。

正如第 2 章描述的，无向概率图模型适合对无法描述变量间方向性的问题进行建模。由于马尔可夫随机场(MRFs)和条件随机场(CRFs)是应用最广泛和最具代表性的无向概率图模型，因此，在本章对两者分别从概念、参数学习方法到结构学习方法作一详细论述。

5.2 马尔可夫随机场

马尔可夫随机场(Markov Random Fields，MRFs)是概率理论的一个分支，用来分析物理现象的空间关系或者非因果上下文关系。在图像处理领域，马尔可夫随机场理论能够很好地描述相邻的图像像素或者相关特征间的相互依赖关系。Hamersley 和 Clifford 确定了马尔可夫随机场与 Gibbs 分布之间的等价性。Besag 进一步将其发展为马尔可夫联合分布。这些研究成果使得可以在贝叶斯框架下用数学方法来建模图像分析中的视觉问题。马尔可夫模型主要用于图像的低层次的处理，如图像复原与分割、表面重建、边缘检测、纹理分析等，而在高层次主要用于目标匹配与识别。

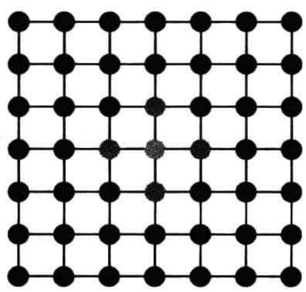


图 5.1 一个马尔可夫随机场示例

假定让 $X=\{X_1,X_2,\cdots,X_N\}$ 表示定义在一组位置(Site)集合 V 上的一组随机变量集合，称为随机场，那么当且仅当满足以下条件时，随机场 X 是在 V 上关于领域系统 N 的马尔可夫随机场，即

$$P(x)>0, \forall x \in X, \quad \text{非负性};$$

$$P(x_i \mid x_{V|\{i\}}) = P(x_i \mid x_{N_i}) \quad \text{马尔可夫性};$$

其中，$x_{N_i}=\{x_{i'} \mid i' \in N_i\}$，$x_i$ 为随机变量 X_i 的一个取值。在图 5.1 中，灰色的结点为 X_i，黑色结点为 X_i 的邻域结点。

5.2.1 邻域系统和团

邻域关系具有以下特性：

① 某结点的邻域不包括结点本身：$i \notin N_i$。

② 邻域关系是相互的：$i \in N_{i'} \Leftrightarrow i' \in N_i$。

③ 邻域系统有不同的阶数。对于二维图像，一阶邻域系统，又称为四邻域系统，如图 5.2(a)所示，每个内部的点都有 4 个相邻点，X 表示被考虑的点，O 表示它的相邻点。2 阶邻域系统，也称为八邻域系统，如图 5.2(b)所示，每个内部的点都有 8 个相邻点。图 5.2(c)中的数字 $n=1,2,\cdots,5$ 表示在 n 阶邻域系统中最外层的相邻点。

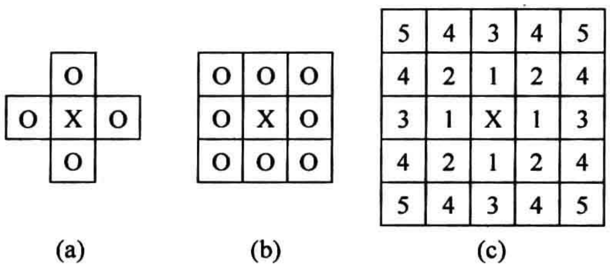


图 5.2 领域关系

一个团(Clique)c 被定义为一组位置集合 V 的一个子集，在一个团里面，所有的结点对都相互连接。例如，对于一个八邻域系统，如图 5.3 所示，图 5.4 给出了它所有的团类型。第一个为空团，第二个为单点团，第三个与第四个分别为水平与竖直方向的两点团，第五个为对角线形的两点团，第六个为三点团，第七个为四点团。可以看出，随着阶数的增长，团的数量也随着急剧上升，同时计算量也增加。

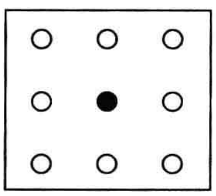


图 5.3 一个八邻域系统

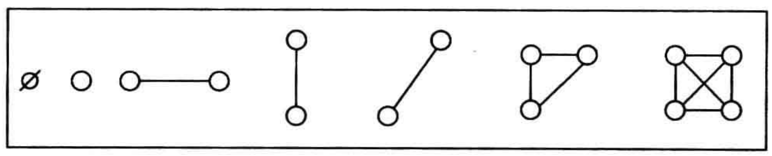


图 5.4 八邻域系统的团类型

当且仅当它不是其他团的子图时，一个团为极大团(Maximal Clique)；当且仅当它的点集模最大，一个团为最大团(Maximum Clique)。如图 5.5 所示，极大团有{2,3}、{3,4}、{4,5}、{4,6}和{1,2,5}，最大团为{1,2,5}。

在图 5.6 中，极大团为{6}、{1,2}、{4,5}和{2,3,4}，最大团为{2,3,4}。

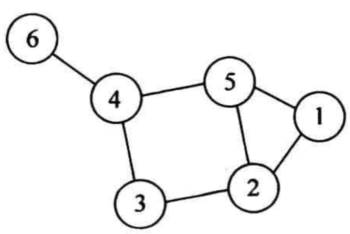


图 5.5 极大团和最大团

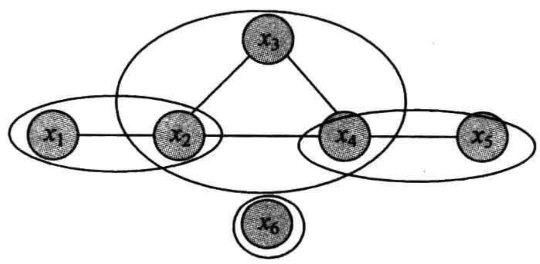


图 5.6 极大团和最大团

5.2.2 HC 定理

给定一个 MRFs，HC(Hammersley-Clifford)定理可以被用来计算联合概率分布，即

$$p(x) = \frac{1}{Z}\prod_{c\in C}\Psi_c(x_c) \tag{5.1}$$

式中,c 为极大团;C 为所有极大团的集合;x_c 为团 c 上的一组随机变量取值;Z 是一个归一化常量,通常称为配分函数,确保 $p(x)$的积分为 1;Ψ_c 为非负势函数。

例如,在图 5.6 中,联合概率分布可以写成式(5.2),即

$$p(x) = \frac{1}{Z}\Psi_{12}(x_1x_2)\Psi_{234}(x_2x_3x_4)\Psi_{45}(x_4x_5)\Psi_6(x_6) \tag{5.2}$$

5.2.3 Pairwise MRF 模型

Pairwise MRF 模型是广泛应用的一种 MRFs 模型,如图 5.7 所示,其结构形式简单和计算量低。

根据 HC 定理,其联合概率分布可以写成以下形式,即

$$p(x,y) = \frac{1}{Z}\prod_{i,j}\Psi(x_i,x_j)\prod_{i}\Phi(x_i,y_i) \tag{5.3}$$

式中,x 为标注类;y 为观测值。

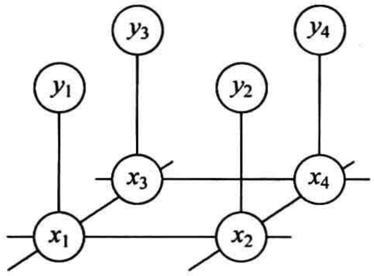


图 5.7 Pairwise MRF 模型

5.2.4 MRFs 的参数学习

1. 最大似然估计法

可以基于最大似然估计法来学习无向概率图模型的参数。然而,由于归一化常数 Z 联合了模型中所有参数,因此无向概率图模型的参数学习更加困难。

例如,给定一组独立同分布的训练数据 $\{x^k\}_{k=1}^K$,想要学习以下的一个 MRF 模型参数,即

$$p(X\mid\theta) = \frac{1}{Z(\theta)}\exp\Big[-\sum_{c\in C}V_c(X,\theta)\Big] \tag{5.4}$$

注意到配分函数 Z 是依赖于参数的,因为它是由式(5.5)计算出来的,即

$$Z = \sum\nolimits_X \exp\Big[-\sum\nolimits_{c\in C}V_c(X,\theta)\Big] \tag{5.5}$$

基于最大似然参数估计法的原理,最优参数可以通过式(5.6)被估计出来,即

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\prod_{k=1}^{K}p(x^k\mid\theta) \tag{5.6}$$

通过使用梯度上升方法来最大化 Log-likehood(对数似然函数),可以获得式(5.6)的最优化参数,对数似然函数为

$$L(\theta) = \sum_{k=1}^{K}\log p(x^k\mid\theta) = \sum_{k=1}^{K}\Big[-\sum_{c\in C}V_c(X,\theta) - \log Z(\theta)\Big] \tag{5.7}$$

对数似然函数的微分方程为

$$\begin{aligned}\frac{\partial L(\theta)}{\partial(\theta)} &= \sum_{k=1}^{K}\Big[-\sum_{c\in C}\frac{\partial V_c(X,\theta)}{\partial\theta}\Big|_{X=x^k} - \frac{\partial\log Z(\theta)}{\partial\theta}\Big] \\ &= \sum_{k=1}^{K}\Big[-\sum_{c\in C}\frac{\partial V_c(X,\theta)}{\partial\theta}\Big|_{X=x^k} + \sum_{X}p(X\mid\theta)\Big(\sum_{c\in C}\frac{\partial V_c(X,\theta)}{\partial\theta}\Big)\Big]\end{aligned} \tag{5.8}$$

在这个公式中,第二个梯度项来自于归一化常数。注意到这个梯度项包含了所有 X 的

概率值的和。当 X 的随机变量数目很大时,计算量是指数增加的,并会迅速导致不可计算。

不过,几个近似的参数学习方法已经被提出来解决这个问题。在最大似然估计框架中,这些方法以不同的方式近似估计似然函数或配分函数 $Z(\theta)$ 的期望值。

2. *近似方法*

MRF 模型参数学习的近似方法包括以下几种:

(1) Iterative Proportional Fitting (IPF) [304]。

(2) Pseudo-likelihood approximation [305]。

(3) Saddle point approximation [306]。

(4) Piecewise training[307]。

(5) Approximate ML estimation by pseudo—moment matching[308]。

(6) Parameter learning with Gibbs/MCMC sampling[309]。

(7) Learning with contrastive divergence[310]。

(8) Max-Margin parameter learning[311]。

这些方法的主要原理是避免精确地计算部分函数的对数及其微分。例如,Pseudo-likelihood approximation 使用局部 Pseudo-likelihood 的乘积近似估计似然,这可以通过假设近邻变量是已知来进行计算。给定这个假设,Pseudo-likelihood 能够被局部归一化,这是很容易计算出来的。这个性质使得 Pseudo-likelihood approximation 成为一个非常有效的参数学习方法,虽然有一定的精度损失。更详细地讨论这些方法,读者可参阅相关的原始论文。

总而言之,最大似然估计方法和近似方法都是基于完整的训练数据来学习 MRFs 参数的,即训练数据的所有类标签都是已知的。对于不完整数据的学习,通常使用 EM 方法,EM 方法首先推理出缺失的数据的类标签(E-step),然后使用整个数据集更新模型的参数(M-step),使用这种方式,EM 把不完整数据集的参数学习转化成了完整数据集的参数学习。EM 迭代进行该过程,直到学习过程收敛。

5.3 条件随机场

条件随机场(Conditional Random Fields, CRFs)的概念由 John Lafferty、Andrew McCallum 和 Femando Peerira 在 2001 年首次提出。自被提出以来,在计算机视觉、自然语言处理等领域得到了广泛的应用。CRFs 是用来标注和划分序列结构数据的概率化结构模型。言下之意,就是给定的输出标识序列 X 和观察序列 Y,条件随机场通过定义条件概率 $p(X|Y)$,而不是联合概率 $p(X,Y)$来描述模型,常见的统计模型可以分为以隐马尔可夫模型为代表的生成模型和以最大熵模型为代表的判别模型。隐马尔可夫模型的假设条件为:某时刻的观测值依赖于该时刻的隐藏状态,同时各个观察值是相互独立的,但实际条件下这一假设通常很难满足。相对于隐马尔可夫模型,CRFs 的主要优点在于它的条件随机特征,只需要考虑当前已经出现的观测状态特征,没有独立性的严格要求。而相对于最大熵模型和其他针对线性序列模型条件,隐马尔可夫模型会出现条件偏置(Label Bias)的问题,CRFs 避免了这个问题,对于整个序列内部的信息和外部观测信息都可以有效地应用。CRFs 具

有最大熵模型的一切优点,两者的关键区别在于,最大熵模型使用每一个当前状态的指数模型来计算给定状态的下一个状态的条件概率,而 CRFs 用单个指数模型来计算给定观察的整个标记序列的联合概率。因此,在不同的状态、不同特征的权重下可以相互交替代换。

CRFs 可以看成是一个在给定输入结点条件下,计算输出结点的条件概率的无向图模型。定义 $G(V,E)$ 为一个无向图,V 为结点集合,E 为无向边集合。$X=\{X_i,i\in V\}$,即 V 中的每个结点对应一个随机变量 X_i。如果以观察序列 Y 为条件,每一个随机变量 X_i 都满足以下特性,即

$$p(X_i \mid Y, X_{V/i}) = p(X_i \mid Y, X_{N_i}) \tag{5.9}$$

式中,N_i 为相邻的结点。那么,(X,Y) 为一个条件随机场。根据 HC 定理,可以得到下面的公式,即

$$p(X \mid Y) = \frac{1}{Z}\prod_i \varphi(X_i, Y, \theta_i) \prod_{i,j\in N_i} \Psi(X_i, X_j, Y, \mu_i) \tag{5.10}$$

理论上,只要在标记序列中描述了一定的条件独立性,G 的图结构可以是任意的,对序列建模可以是最简单的、最普通的链式(Chain Structured)图结构,结点对应标识序列 X 中的相应元素,如图 5.8 所示。

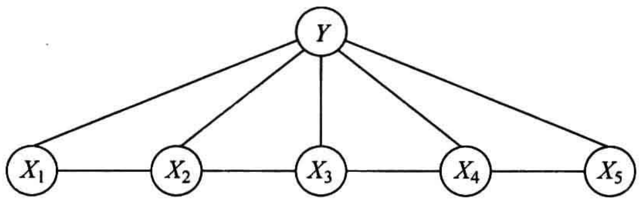


图 5.8 一个 CRFs 模型

假设图 5.8 中的结点满足 $p(X_i|Y,X_{V/i})=p(X_i|Y,X_{N_i})$,则

$$p(X_3 \mid Y, X_{V/3}) = p(X_3 \mid Y, X_2, X_4) \tag{5.11}$$

$$p(X_1, X_2 X_3, X_4 \mid Y) = \frac{1}{Z}\varphi(X_1, Y, \theta_1)\Psi(X_1, X_2, Y, \mu_1)\cdots\varphi(X_5, Y, \theta_5)\Psi(X_5, X_4, Y, \mu_5) \tag{5.12}$$

从图 5.8 中还可以看出,观察序列 Y 的元素之间不存在图结构,因为这里只是将观察序列作为条件输入,并不对观察序列 Y 做任何假设。

条件随机场模型需要解决 3 个基本问题:特征的选取、参数训练和解码。其中参数训练的过程可在训练数据集上基于对似然函数的最大化进行。

对于一个定义为形式如式(5.13)的 CRFs 模型,有

$$p(X \mid Y) = \frac{1}{Z}\exp\Big(\sum_{i,k}\lambda_k f_k(x_{i-1}, x_i, y) + \sum_{i,k}\alpha_k g_k(x_i, y)\Big) \tag{5.13}$$

其中,

$$Z(Y) = \sum_y \exp\Big(\sum_{i,k}\lambda_k f_k(x_{i-1}, x_i, y) + \sum_{i,k}\alpha_k g_k(x_i, y)\Big) \tag{5.14}$$

其中,每个 $f_k(x_{i-1},x_i,y)$ 是标识序列 x 中位置为 i 和 $i-1$ 的输出结点的特征函数,每个 $g_k(x_i,y)$ 是位置为 i 的输入结点和输出结点的特征函数,每个 λ 和 μ 是特征函数的权重,可以从训练数据中估计得到。通过分别为训练数据中的每一个状态—状态 (x',x) 和状态—观测 (x,y) 定义特征函数,使得建立的 CRFs 模型具有类似最大熵模型的特征,即

$$f_{x',x}(x_u,x_v,y)=\begin{cases}1, & \text{if} x_u = x',x_v = x\\ 0, & \text{otheruise}\end{cases} \tag{5.15}$$

$$g_{x,y}(x_v,y)=\begin{cases}1, & \text{if} x_v = x,y_v = y\\ 0, & \text{otheruise}\end{cases} \tag{5.16}$$

可以用广义迭代算法(Generalize Iterative Scaling,GIS)对 CRFs 模型进行训练。

如同前面几种模型一样,CRFs 模型中的参数也需要通过训练得到。在 CRFs 中参数是指式(5.13)中的 λ,它是对应于每个特征函数的权重。

5.3.1 问题分析

贝叶斯学派认为,所有的未知属性都应当看作随机变量,特定参数下 λ 的训练数据 T 概率看作是条件概率分布 $p(T|\lambda)$。和他们不同的是频率论学派并不把 λ 看作是随机变量。相反,特定 λ 下的训练数据的概率被当作训练数据上的一个用 λ 标记的分布族,写成 $p_\lambda(T)$ 或者就记为 $p(T|\lambda)$。频率论方法进行参数估计需要借助估计器(Estimator)的使用,在频率论体系中最为常用的估计器就是极大似然估计器。因此,CRFs 模型的最大后验概率训练可以用最大似然的方法,也就是在给定输入序列的情况下估计模型中相应的参数,使它们对应的标注序列构成的集合的条件概率最大化。

假设训练集上的数据 $T=\{(x_k,y_k)\}(\forall k=1,\cdots,N;N$ 为训练数据集的规模)独立且与联合经验分布 $\tilde{p}(x,y)$ 同分布。应用于条件模型 $p(x|y,\lambda)$ 得到训练数据的相似度函数为

$$L(\lambda)=\prod_{x,y}p\ (x\mid y,\lambda)^{\tilde{p}(x,y)} \tag{5.17}$$

最大似然估计(MLE)利用似然函数选择可能的参数 λ,能够使相似度函数最大化的参数也使得模型分布更接近于经验分布,MLE 原理指出能够满足相似度函数最大化的参数 λ 值就是所需要的参数值,即

$$\lambda_{\mathrm{ML}}=\arg\max L(\lambda) \tag{5.18}$$

为了运算方便,相似度函数式(5.17)往往表示成对数形式,因此式(5.17)又可以写成

$$L(\lambda)=\sum_{x,y}\tilde{p}(x,y)\log p(x\mid y,\lambda) \tag{5.19}$$

CRFs 模型中,训练集上所有序列 $p(x^{(k)}\mid y^{(k)},\lambda)$ 的乘积就是相似度。对于 CRFs 模型利用对数极大似然估计方法能够得到使相似度最大化的参数值。其对数相似度形式为

$$L(\lambda)=\sum_{k}\left[\log\frac{1}{Z(y^{(k)})}+\sum_{j}\lambda_j f_j(x^{(k)},y^{(k)})\right] \tag{5.20}$$

这个函数是凸函数,从而保证了函数会收敛于全局最大值(Global Maximum)。这个相似度函数对于 λ_i 的一阶导数形式为

$$\begin{aligned}\frac{\partial L(\lambda)}{\partial\lambda_k}&=\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n}f_k(x_{i-1},x_i,y)-\sum_{x,y}\tilde{p}(y)p(x\mid y,\lambda)\sum_{i=1}^{n+1}f_k(x_{i-1},x_i,y)\\&=E_{\tilde{p}(x,y)}[f_k]-E_{p(x|y,\lambda)}[f_k]\end{aligned} \tag{5.21}$$

式中,$\tilde{p}(x,y)$ 为训练数据的经验分布;$E_p[.]$为分布 p 的期望。将这个一阶导数取零,就是最大熵模型的约束条件:与模型分布相对应的每个特征的期望等于经验分布下的期望。可以看出 CRFs 模型符合最大熵原理。

式(5.21)的经验分布 $E_{\tilde{p}(x,y)}[f_j(x,y)] = \left(\sum_{k=1}^{N} F_j(x^k,y^k)\right)$，其中 $F_j(x^k,y^k)$ 表示给定观察序列 y 和标注序列 x 的特征 f_j 的个数，即 $\sum_{i=1}^{n} f_j(x_{i-1},x_i,y,i)$。

对式(5.21)，通过常规的方法求取斜率为零的情况来估计参数值 λ 很困难，而且一般无法得到一个封闭的解，这就涉及训练算法的选择问题。最大熵模型使用的训练算法，如GIS、IIS等，都可以应用于CRFs模型的训练。另外，变梯度方法、爬山方法也可以。Hessian矩阵的逆矩阵的计算，有更高的收敛速度。下面将对CRFs模型参数估计中的训练算法进行介绍。

5.3.2 模型训练中的动态规划

在对模型中的参数进行估计时，无论采取迭代收敛还是基于梯度的训练算法，对于训练数据中的每个观测序列 $x^{(k)}$，都需要有效地计算每个特征函数在模型分布下的期望值 $E_{p(x|y^{(k)},\lambda)}[f_j(x,y^{(k)})]$，它又可以表示为

$$\begin{aligned} E_{p(X|y^{(k)},\lambda)}[F_j(X,y^{(k)})] &= \sum_{x} p(X = x \mid y^{(k)},\lambda)F_j(x,y^{(k)}) \\ &= \frac{1}{Z(Y^k)}\sum_{X'}\left[\exp\left(\sum_j \lambda_j F_j(X',Y^k)\right)\times \sum_j F_j(X',Y^k)\right] \end{aligned} \tag{5.22}$$

式(5.22)的计算量是十分巨大的，如果观察序列 $y^{(k)}$ 具有 n 个元素，那么相对应的标注序列就有 $n^{|X|}$ 种可能，在这个数量级上求和的时空开销太大，只能改用别的方法。式(5.22)中的右边部分可以写成

$$\sum_{i=1}^{n}\sum_{x',x} p(X_{i-1} = x', X = x \mid y^{(k)},\lambda) f_j(x',x,y^{(k)}) \tag{5.23}$$

这样就不需要在所有 $n^{|X|}$ 个序列上计算。利用HMM模型中的前向后向(Forward-Backward)算法的动态规划方法可以计算 $p(X_{i-1}=x',X=x|y^{(k)},\lambda)$。

分别定义一个前向向量 $\boldsymbol{\alpha}_i(y)$ 和一个后向向量 $\boldsymbol{\beta}_i(y)$，即

$$\boldsymbol{\alpha}_0(x \mid y) = \begin{cases} 1 & \text{如果 } y = \text{开始状态} \\ 0 & \text{其他} \end{cases}$$

$$\boldsymbol{\beta}_{n+1}(x \mid y) = \begin{cases} 1 & \text{如果 } y = \text{停止状态} \\ 0 & \text{其他} \end{cases}$$

则有以下的递推关系，即

$$\boldsymbol{\alpha}_i(y)^T = \boldsymbol{\alpha}_{i-1}(y)^T M_i(y)$$

$$\boldsymbol{\beta}_i(y)^T = M_{i+1}(y)\boldsymbol{\beta}_{i+1}(y)$$

其中 $M_i(y)$ 是随机变量矩阵，$M_i(y)$ 中的每个元素 $M_i(x_{i-1},x_i|y)$ 定义为：

$$M_i(x_{i-1} = x_i' y_i = x \mid y) = \exp\left(\sum_k \lambda k f_k(x_{i-1},x_i,y,i)\right)$$

Y_i 和 Y_{i-1}，在给定观察序列 $y^{(k)}$ 下取得标记 x' 和 x 的概率为

$$p(X_{i-1} = x', X_i = x \mid y^{(k)},\lambda) = \frac{\boldsymbol{\alpha}_{i-1}(x' \mid y)M_i(x',x \mid y)\boldsymbol{\beta}_i(x \mid y)}{Z(y)} \tag{5.24}$$

$Z(x)$ 通过 $n+1$ 个 $M_i(x)$ 矩阵的(start,stop)部分的乘积得到，可以表示为

$$Z(y) = \boldsymbol{\alpha}_{n+1}(y^k) = \boldsymbol{\beta}_0(y^k) \tag{5.25}$$

将其代入式(5.24)中,就可以推导出计算特征期望的动态规划方法。

5.3.3 参数估计的训练算法

CRFs 模型训练过程分为两部分,第一部分为特征函数集的建立,第二部分为模型参数估计,即计算特征函数的权重。前面已经对第一部分的工作进行了详细的介绍,下面主要介绍几种常用的训练算法。

通常迭代梯度方法(Iterative Scaling)用于 CRFs 的训练。GIS(Generalized Iterative Scaling)算法和 IIS(Improved Iterative Scaling)算法都属于这种训练方法的范畴。迭代梯度方法比较简单、容易实现,但是存在收敛慢的缺点。Wallach 尝试了将变化斜率(Conjugate Gradient)方法和 2 阶方法用于 CRFs 训练,达到了比较好的效果。L-BFGS 算法相对收敛速度比较快,它是一种近似的 2 阶方法。不管是迭代梯度方法还是近似 2 阶方法都是通过似然函数最大化来得到参数的。下面介绍一下这几种常用的训练算法。

1. GIS 算法

GIS 算法是由 Darroch 和 Ratliff 于 1972 年提出的,可以用于求解特征向量。设有 K 个特征函数 f,则需估计的参数向量为$\{\lambda_1\lambda_2\cdots\lambda_k\}$。

首先,GIS 算法要求,对于任何 $y,x\in Y\times X$,K 个特征函数之和为常量,即

$$\sum_{j=1}^{k} f_j(y,x) = C \tag{5.26}$$

如果这一条件在实际使用中不能得到满足,则引入一个新的特征函数 $f_l(l=k+1)$,有

$$f_l(y,x) = C' - \sum_{j=1}^{k} f_j(y,x) \tag{5.27}$$

其中,$C' = \max\limits_{x\in X, y\in Y} \sum_{j=1}^{k} f_j(y,x)$。

注意,$f_l(y,x)$和普通的特征函数不同,它的变化范围是 $0\sim C'$。

下面介绍 GIS 算法。可以证明该算法收敛于 P^*。

$$\lambda_j^{(0)} = 1$$

$$\lambda_j^{(n+1)} = \lambda_j^{(n)} \left[\frac{E_{\bar{p}} f_j}{E_{p^{(n)}} f_j}\right]^{\frac{1}{C}}$$

其中,$E_{p^{(n)}} f_j = \sum_{x,y} \bar{p}(x) p^{(n)}(y|x) f_j(y,x)$,$p^{(n)}(y|x) = \frac{1}{Z(b)} \prod_{j=1}^{l} (\alpha_j^{(n)})^{f_j(y,x)}$。Darroch 和 Ratcliff 证明了,$L(P)$单调递增,也就是说 $L(P^{(n+1)}) \geqslant L(P^{(n)})$,且$\lim_{n\to\infty} P^{(n)} = P^*$。

对于 GIS 算法,有以下步骤:

(1) $E_{\bar{p}} f_j$ 的计算,即

$$E_{\bar{p}} f_j = \sum_{x,y} \bar{p}(y,x) f_j(y,x) = \frac{1}{N} \sum_{i=1}^{N} f_j(y_i,x_i) \tag{5.28}$$

式中,N 为训练样本特征函数的个数。在整个递归迭代过程中只需要计算一次。

(2) $E_{p^{(n)}} f_j$ 的计算。在每次迭代过程中,需计算 $E_{p^{(n)}} f_j = \sum_{x,y} \bar{p}(a) p^n(y|a) f_j(y,x)$。它是训练过程中最耗时的操作。

(3) 终止条件。当 $L(p(n+1))-L(p(n))<\varepsilon$，或者精度的增长已不明显，就可考虑终止迭代，因为这时，更多的迭代不会使性能有明显的提高。

2. IIS 算法

20 世纪 80 年代，Della Pietra 兄弟在 IBM 对 GIS 算法进行了两方面的改进，提出了改进迭代算法 IIS(Improved Iterative Scaling)。这使得模型的训练时间缩短了一到两个数量级。IIS 算法适用于任何特征函数非负的情况。算法描述如下：

Algorithm5.1 IIS algorithm

$\{\lambda_1\lambda_2\cdots\lambda_n\}=\text{IIS}(f_1,f_2,\cdots,f_n,\bar{p}(x,y))$

//$f_1,f_2,\cdots,f_n$ 为 n 个特征函数，$\bar{p}(x,y)$为经验概率分布；

//$\lambda_1\lambda_2\cdots\lambda_n$ 为 n 个参数值；

1 初始化 $\lambda_i=0, \forall i\in\{1,2,\cdots,n\}$

2 对 $i\in\{1,2,\cdots,n\}$，

a. 令 $\Delta\lambda_i$ 为

$$\sum_{x,y}\bar{p}(x)p(y\mid x)f_i(x,y)\exp(\Delta\lambda_i f^n(x,y))=\bar{p}(f_i)$$

的解，其中，$f^n(x,y)\equiv\sum_{i=1}^{n}f_i(x,y)$.

b. 用 $\lambda_i\leftarrow\lambda_i+\Delta\lambda_i$ 更新 λ_i 的值.

3 如果还有 λ_i 不收敛，转步骤 2；否则，算法结束.

3. L-BFGS 算法

前面曾经提到，CRFs 模型符合最大熵原理，因此，迭代梯度算法可以用于 CRFs 模型的训练。虽然 IIS 算法相对于 GIS 在训练时间上大为缩短，但是仍然比较长。在这里介绍另一种近似的 2 阶方法 L-BFGS 算法，利用这种方法训练 CRFs 模型比迭代梯度方法快很多，同时比变化斜率方法也要快。

对于非线性优化问题，牛顿方法利用二阶信息也就是利用曲率信息确定搜索方向。作为一种 Quasi-Newton 方法，L-BFGS 通过之前斜率和增量来估计曲率，通过保留针对函数一阶导数的运动的有限尺寸窗口，近似地得到相似度的 2 阶导数。

L-BFGS 算法可以看成黑盒优化过程，外界只需提供需要优化的函数的 1 阶导数即可。式(5.21)中已经给出了似然函数的一阶导函数形式，为了克服数据稀疏问题，式(5.21)可以进行以下变化，即

$$\frac{\partial L(\lambda)}{\partial\lambda_j}=E_{\bar{p}(y,x)}[f_j(y,x)]-\sum_k E_{p(y|x^{(k)},\lambda)}[f_j(y,x^{(k)})]-\frac{\lambda_k}{\sigma^2}\tag{5.29}$$

注意到，式(5.29)和式(5.21)的区别在于式(5.29)多了一个因子$\frac{\lambda_k}{\sigma^2}$，该因子在这里起到一种平滑作用。由于数据样本的规模和它包含的数据类别有限，从而导致数据稀疏现象的产生。数据稀疏现象是指，在基于统计技术的知识获取方法中，在数据样本的规模不够大的条件下，某种属性变量的关系在数据样本中出现的次数很少，甚至根本不出现的现象。即许多合法的、在未来的数据中可能要遇到的属性关系在统计数据样本中出现的次数很少或

从未出现过，从而造成知识短缺的现象。在实际应用中，数据稀疏的存在会产生大量空值，严重影响后续处理的性能和效果。而训练数据规模的扩大可能付出的代价是相当大的。参数平滑就是在训练数据不足够充分的条件下，采用某种方式对统计结果和概率估计进行必要的调整和修补，以降低由于数据稀疏现象带来的统计误差。在这里因子 σ 就起到了这种作用。

式(5.29)前两项为特征 f_j 的经验期望值和模型期望之差，和迭代梯度方法类似，面临着求解特征函数的模型期望这个问题。通过 CRFs 模型的矩阵计算和动态规划方法，这个问题可以很好地解决。

5.3.4 参数估计的训练过程

下面介绍一下 CRFs 模型训练的过程。

模型的训练过程描述如下：

输入　特征模板集，迭代次数，特征出现次数，平滑因子 σ。

输出　特征函数集 F 及其参数集 λ：

训练流程如下：

(1) 通过动态规划方法获得特征函数的模型期望和经验期望。

(2) 通过平滑因子 σ，特征的模型期望，经验期望得到 $\frac{\partial L(\lambda)}{\partial \lambda_j}$，送入训练算法模块进行训练并且获得修正后的参数。如果到了迭代中止条件、退出循环、训练结束，否则转到(1)。

模型训练模块结构如图 5.9 所示。

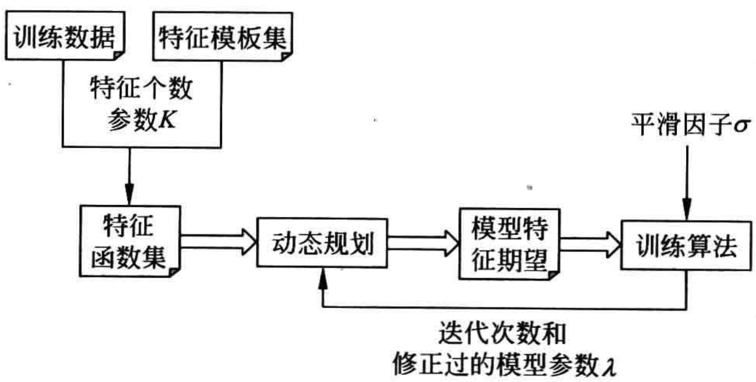


图 5.9　CRFs 的训练过程

平滑因子 σ 能够解决训练规模不足造成的数据稀疏问题。λ 的最终收敛程度由迭代次数决定。

第6章 概率图模型的新型学习方法

6.1 概述

近年来，随着概率图模型应用领域的不断扩大，在模式识别和数据挖掘等研究任务中，传统的概率图模型的学习方法遇到了新的挑战。例如，系统建模所需的大量带类标签训练数据获取十分困难，应用系统要求在线学习，或者由于计算系统内存不足导致大数据量的训练样本不能一次性装载等特殊问题。这些问题的出现，导致了传统的概率图模型学习方法不能进行应用，或者所建模型性能较差。

为此，研究者们在基于传统学习方法的基础上，提出了多种概率图模型的新型学习方法。本章主要针对贝叶斯网络的主动学习和增量学习方法进行详细论述。

6.2 主动学习方法

主动学习的目标是选择较少有价值样本数据学习高质量分类或识别模型。将其引入概率图模型的学习，主要是解决大数据量带类标注数据难以获得的问题。

机器学习实质上是根据特殊实例得到一般概念，并将此概念运用到未知样本的归纳及演绎过程。传统机器学习中的训练样本通常由外部施教者提供，或任意采样自潜在的数据分布，也称为被动学习(Passive Learning)。而假若学习器能够主宰学习进程，可自主选择需要查询(或标记)的实例，然后基于外部施教者对实例的真实响应(或分类)进行学习，则会显著改善学习过程和结果，此为主动学习(Active Learning)。

相对于传统被动学习，主动学习中学习器取得一定主动权通过研究数据聚集类或分布结构等，选择对学习过程最有用的代表性样本请求用户响应(或标记)，以加快学习过程并提高所学概念的准确性，同时减少所需训练的样本数目。主动学习器选择实例并请求外界响应的过程称为查询(Query)。

主动学习方法主要研究所应采取的选择策略以提出好的查询，其寻找能最有效利用有限训练样本的途径，通过尽量少的查询样本获取最大限度的有用信息。本章详细介绍主动学习的原理及其常用选择策略，并对基于主动学习的贝叶斯网络分类器的建立和贝叶斯网络结构的主动学习方法进行阐述。

6.2.1 主动学习原理

1. 基本原理

传统监督学习通常利用任意采样的方式从潜在分布中收集大量样本，然后根据样本推断分类器或模型假设，可称为被动学习，如图 6.1 所示。被动学习器是“从样本中学习”，它随机地选择训练样本，被动地接受这些样本的信息，输出分类器或模型假设。目前大部分模式识别算法（如 SVM、神经网络、贝叶斯网络等）几乎都采用被动学习方式，它使用给定的大数据量训练样本学习模型参数和结构，其训练数据必须完整。

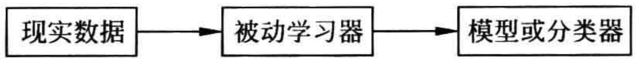


图 6.1 被动学习方式

然而样本的收集与标注要花费巨大的人力和物力，在整个监督学习过程中所占比例更是达到令人吃惊的地步。因此，学习器所得到的训练样本集比较有限，导致纯粹的监督学习算法泛化能力较低。而正是由于用户标记的样本数目有限以致训练数据的获取需付出较大代价，使得如何选择更为有效的样本成为一个极具价值的研究问题。

一般学习器通常在假设数据为独立同分布情况下，通过任意取样得到训练样本集，但所得到的训练样本集实际是有冗余的。如果对采样过程进行合理引导，可利用较少训练样本达到相同学习效果。具体方法则是考虑赋予学习器对于样本的选择权主动查询用户，也即为主动学习方法的基本思想。

主动学习基于对数据分布的分析及外界对已有样本的响应（标记）情况，通过对特定样本的查询来引导采样过程，如图 6.2 所示，也称为查询方式的主动学习算法。主动学习器在查询外界施教者及接收反馈的交互过程中获取信息，输出分类器或模型假设。学习的初始阶段，多数样本有用性程度较高，随着交互过程的进行，模型性能得到提高，此时，仅有少部分样本对于模型性能的进一步提升是有用的，通过样本的合理选择，完全有可能减少所需训练样本的数量。

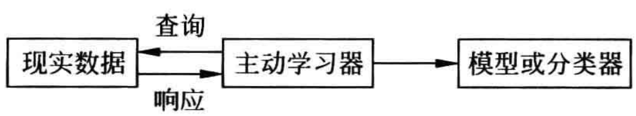


图 6.2 主动学习方式

主动学习与被动学习的不同在于，被动学习只是简单接收外界的任意信息，而主动学习可主动选择所期望的信息，其性能在很多情况下相对被动学习有所提高。主动学习的研究目标是寻找某种途径来选择对于学习过程最有用的查询（样本），以在查询过程中得到尽可能多信息的同时，利用所得信息尽快终止搜索过程。最有用查询（样本）则是指能够最大可能改善当前所得模型性能的查询（样本），以减少所需响应的查询数目，同时最小化查询学习所需的迭代次数以加快学习过程及提高所学概念的准确性。Freund 等从理论上证明，在一定假设条件下，采用合理的主动采样策略达到相同学习效果（即相同泛化误差，Generalization Error），所需训练样本可减少到任意采样情况下的对数倍，而假设通过查询用户总是可以得到高的期望信息，主动学习的泛化误差将以所查询样本数量指数倍的速度

下降。

主动学习的目的是使用尽可能少的训练数据来学习高性能的分类器，从而有效减少人工标注样本的代价。主动学习目前有相当多的研究，其基本原理是基于少量的完整训练样本 L 建立初始模型(或分类器)，每次学习过程中学习器可以主动在不完整(或未带类标注的)候选样本集 U 中选择最有利于模型性能的样本 X_i，并将这些样本以一定的方式加入到训练集中来进一步训练模型，如图 6.3 所示。

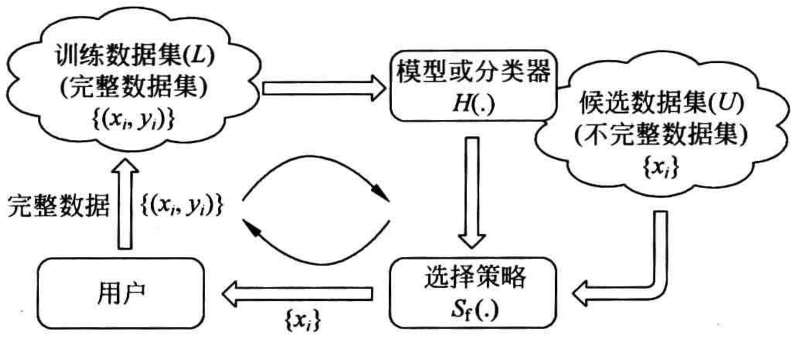


图 6.3 主动学习的基本原理

有两种主动学习模型：一种为基于池(Pool-based)的主动学习模型；另一种为基于流(Stream-based)的主动学习模型(见图 6.4)。

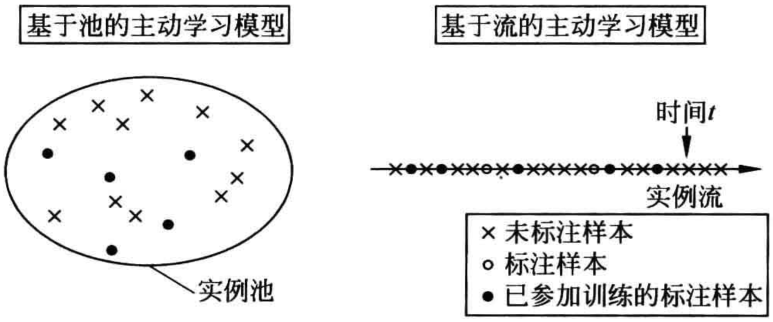


图 6.4 基于池与基于流的主动学习模型比较

基于流的主动学习模型在每次选择当前输入的一个查询(或未标记的样本)，然后按照某种启发式规则决定是否选用该查询(样本)。如果该查询(样本)被选择，则向外界施教者询问其实际的响应(或类标记)，将它加入训练集，重新训练模型(或分类器)，接着利用当前的模型(或分类器)再选择下一个查询(或未标记的样本)继续学习，直到查询集(或未标记样本集)为空或达到某种精度。

1994 年由 Lewis 和 Gale 提出了基于池的主动学习模型，在该模型中，由所有查询(或未加标记的样本)组成一个样本池，而学习器在学习过程中可以访问这个池，并可以询问池中任意一个查询(或样本)的实际响应(或标记)。其学习也是一个循环反复的过程。首先，主动学习器从很少的完整样本开始学习训练一个分类器，然后按照某种启发式规则(采样算法)，选择另外的很少的被认为是最有利模型性能的查询(或样本)，询问它们的实际响应(或标记)，将它们加入到原来的训练样本集合中，通过对这个新的训练样本集合的学习，更新原来的知识，接着再选择一些查询(或未标记的样本)继续学习，直到池中样本集为空或达到某种指标。

基于池的主动学习模型有一定的理论缺陷,尤其是不考虑输入数据的分布情况,同时在计算上也有不利的一面,即在选择训练样本数据之前必须搜索整个空间,检测大量的样本数据,才能确定哪些查询(或样本)被选择。

2. 选择策略

构造主动学习器,采取何种选择策略是关键,即如何衡量某查询(或样本)的响应(或标记)所能提供的信息量。评价某个查询(或样本)的方法依赖于当前学习器对未知关系 $p(x,y)$所得到的假设(Hypothesis)或信任(Belief)。通常所选择的查询(或样本)的响应(或标记)需能提供足够信息,以构建使得某精度标准最大化的假设。评价方法直接关系到整个算法的性能。下面以建立分类器为目标,给出几种常用的评价查询(或样本)的采样策略。

(1) 基于不确定性抽样的选择策略

在《机器学习》一书中提到,“最有分类歧义性的实例也一定最能提供新的分类信息”,基于不确定性抽样的选择策略就是选择当前分类器对其类别最不确定的样本查询用户,也称为单学习器的方法。该方法最初来源于 Cohn 等提出的选择性采样方法;Lewis 等提出一种基于池的不确定性采样方法用于文本分类,利用概率分类器选择后验概率 $p(y|\boldsymbol{x})$为 $0.5\pm b/2$ 的样本查询用户;Tong 等提出一种基于支撑向量机的不确定性采样方法(Active SVM),选择距离 SVM 分界面最近的样本来查询用户。基于单学习器的不确定性采样方法利用指定学习器的启发式搜索实现样本选择,其假设前提为当前学习器对于新的样本能够正确分类。该类方法需要分类器不仅提供样本类别,还要给出类别可信度的衡量方法以引导样本的选择过程,统计模糊、最近邻方法及神经网络等都可用于此类主动学习中。基于单学习器的采样方法存在几方面理论上的缺陷,包括真实不确定性的潜在估计及分类器的模型偏置等。

(2) 基于误差减少的选择策略

基于误差减少的选择策略(Error Reduction Sampling,ERS)认为最优的主动学习应该选择能够最大限度地减少分类器在测试集上的分类误差的样本来进行标注,此种方法属于确定性抽样方法。这里用 $X=<X_1,X_2,\cdots,X_m>$表示特征变量集,特征的取值用小写字母 x_i 表示,任一未标注样本 $\boldsymbol{x}=<x_1,x_2,\cdots,x_m>$由各特征变量的取值组成,$Y=\{y_1,y_2,\cdots,y_l\}$代表类别变量,$S$ 表示所选择的未标注样本,R_s 代表分类误差损失的减少,$L=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_N,y_N)\}$为带有类别标签的训练数据集,$L^+=L\cup S$ 代表新的训练集,U 为整个样本数据集,包括已标注和未标注样本,E_L 为期望误差,那么,有

$$E_L=\int_X E_{Y|X}[C(H_L(\boldsymbol{x}),y)]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{6.1}$$

$$R_s=\int_X (E_{Y|X}[C(H_L(\boldsymbol{x}),y)]-E_{Y|X}[C(H_{L^+}(\boldsymbol{x}),y)])p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{6.2}$$

$$S=s_f(U)=\arg\max_S R_s \tag{6.3}$$

式中,$H(x)$为基于已标注样本集学习得到的分类器;$s_f(.)$为选择函数;$C(.)$为误差损失函数,$C(.)$一般有两种形式。

对数损失,即

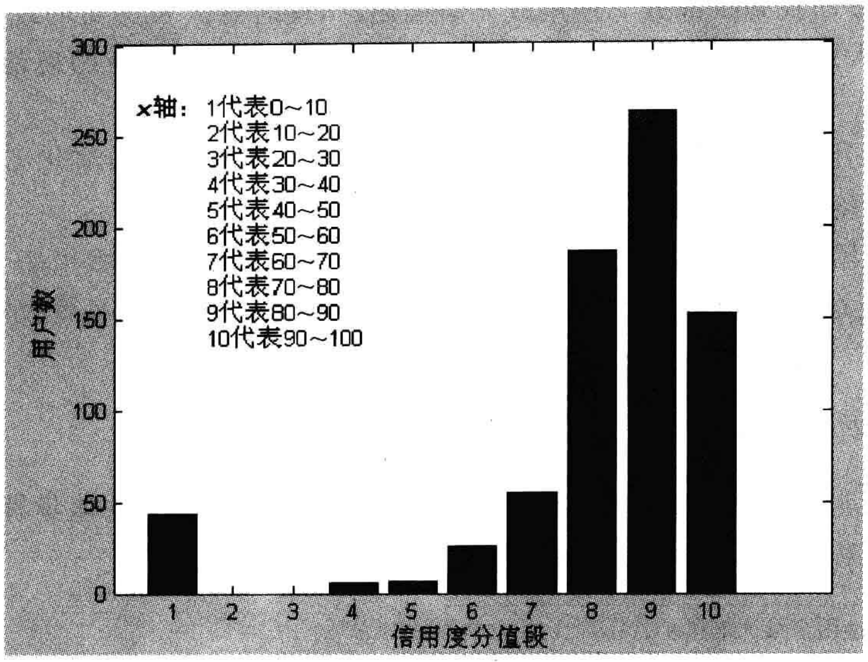


图 8.6 客户信用度分布

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	3	1	2	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	3	1	1	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	3	1	2	2	1	1	2	2	3	1	1	5	2	2	2	50.378
1	1	3	1	2	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	1	2	1	1	1	1	3	1	1	5	2	2	2	82.364
1	1	3	1	2	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	1	2	1	1	1	1	1	1	1	4	2	2	2	90.606
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	5	1	2	2	1	1	1	1	3	1	1	5	2	2	2	82.364
2	2	2	1	2	2	1	2	1	1	3	1	1	4	2	1	1	36.966
1	1	3	2	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	1	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	1	2	1	1	1	1	1	1	3	2	2	2	2	100
1	1	2	1	2	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	2	1	1	2	1	1	1	1	1	1	2	4	2	2	2	79.945
1	1	3	1	1	2	1	1	1	1	3	1	2	3	2	2	2	73.266
1	1	2	1	1	2	1	1	1	1	1	1	1	5	2	1	2	100
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	2	2	1	2	1	1	3	1	1	4	2	2	2	77.08
1	2	3	1	1	2	1	1	1	1	1	1	1	3	2	2	2	91.348
1	1	2	1	2	2	1	1	1	1	3	2	1	3	2	2	2	100
1	1	2	1	1	2	1	1	1	1	1	1	1	4	2	1	2	90.606

图 8.7 信用度评分模型预测结果

8.2.3 基于投票熵和最小相对熵的主动贝叶斯网络建立客户信用等级评估模型

1. 问题背景

客户信用风险等级是电信企业决策支持系统中进行客户总量分析、客户新增与流失分析、客户行为分析、经营风险分析和客户欺诈分析的一个重要分析角度，其可以应用于各种信用风险管理和控制。准确地评估客户信用等级，可以解决在复杂的用户消费群体中良莠

从测试样本中选择出类条件熵最大和最小的候选样本(MinExample,MaxExample),将这两个样本同时加入到训练集中。类条件熵最大的样本的加入,使得分类器能够对具有特殊信息样本的及早重视;而类条件熵最小的样本是分类器较为确定的样本,对它的分类也更加准确,从而部分地抑制了由于不确定性样本的加入而产生的误差传播问题。

宫秀军使用该选择策略对贝叶斯网络参数进行学习,应用到文档分类中,试验结果显示这种方法在处理类别分布均匀时,利用较少带有类别标注的样本获得了与带有大量类别的训练样本几乎相当的精度,但在处理类别分布不均匀时仍然会使选择的测试样本的类别偏向于概率最大的类别,而降低了分类精度。

(4) 基于分类损失与不确定性抽样相结合的选择策略

针对最大最小熵的选择策略的缺陷,宫秀军又提出了基于分类损失与不确定性抽样相结合的选择方法。该方法从测试样本中选择 k 个熵较大的样本,组成集合 $\max S$,然后对此集合中每个元素计算相对于该集合的分类损失和,选择分类损失和最小的样本做标注并加入到训练样本集中。对应于两种分类损失的计算,其相应的计算公式如下。

对数损失,有

$$C = \frac{1}{|\max S|} \sum_{x \in \max S} \sum_{i} p_{D^*}(\hat{y}_i \mid \boldsymbol{x}) \log p_{D^*}(\hat{y}_i \mid \boldsymbol{x}) \tag{6.10}$$

0-1 损失,有

$$C = \frac{1}{|\max S|} \sum_{x \in \max S} (1 - \max_{i}\{p_{D^*}(\hat{y}_i \mid \boldsymbol{x})\}) \tag{6.11}$$

这种方法在类别分布差别较大时效果很好,但是在未带类别标注样本集非常大时,该方法的计算复杂度是相当高的。由于方法(3)和方法(4)是基于池的主动学习方法,每次选择样本都需要检测整个数据集,因而计算复杂度高,在实际应用中难以采纳。

(5) 基于委员会投票的选择策略

基于委员会投票的选择(Query-by-Committee, QBC)也称为多学习器方法。它并不具体地确定搜索空间的大小,而是首先根据假设空间的先验分布,对一系列分类器进行采样,基于分类器组类别预测不一致的程度来查询样本真实标记。这种方法将使得分类界面附近的样本点(也即最富信息量的样本点),而不是位于先验概率最高的区域内样本点,成为训练样本的概率提高。Seung 首先提出了一种通用 QBC 算法,并在理论上证明了随着查询数目趋向于无限,通过 QBC 算法获得的样本信息增益趋向于有限值。Argamon-Engelson(1999)在自然语言处理中使用了 QBC 的方法,其实验证明两个成员的 QBC 的表现同多个成员的 QBC 的表现相当,并且更容易实现,无需参数调整及计算效率更高;同时证明基于流的 QBC 算法要好于基于池的 QBC 算法。基于流的 QBC 主动学习算法伪代码如下:

Algorithm6.1 The QBC of active learning

```
QBC(
A,      //分类算法
K,      //委员会成员数目
L,      //少量带有类别标注的样本集
UL,     //未带类别标注的候选样本集
ζ,      //选择停止的条件
θ,      //投票差异程度的阈值
```

```
)
1 While not ζ{
2   Train K classifiers {M_i: 1≤i≤K} from L based on A      //从 L 中学习 K 个分类器
3     For each x_i ∈ UL
4         For h=1, …, K
5             Compute P(Y|x_i, M_h) //基于 M_h 对x_i 进行分类预测，给出类标注 Y 的分布
6         End For
7         Compute D(x_i) //计算 D(x_i)度量委员会投票差异程度
8         If D(x_i) > θ
9             Select x_i from UL to user for labeling        //从 UL 中选择x_i，交给用户去标注
10          Get the true label y_i of x_i and add (x_i, y_i) into L
                                                             //获得x_i 的真实类别，并将其加入到 L 中
11      End for
12  Train classifier M from L                                //从 L 中训练分类器 M
13  Test M on test data and get its accuracy                 // 测试分类器 M，得到它的分类精度
14 }
15 Return M
```

QBC 选择策略是基于如何减小版本空间(Version Space)，这种策略计算起来比较简单(当评价每个未标注样本时，只需要一次内积运算)，同时该算法倾向于选择的样本数据能够把变形空间分成两个大小近似的部分，这样的样本被加入到训练集以后，其中一个部分被从整个变形空间中除去，最大可能的减小变形空间的期望面积，加快了学习的过程。

对于分类器组的设计，Abe 等使用 Bagging 方法，将可得训练数据集作不同划分，然后根据所划分的不同数据子集训练得到一组分类器；McCallum 等使用 EM 算法来构建委员会成员。

对于委员会投票差异的度量方法目前有两种，即基于相对熵的度量和基于投票熵的度量，两种方法都存在一定的缺陷，导致分类器的分类精度达不到与被动学习相同的精度；并且对于带有噪声的训练样本集，QBC 方法同不确定抽样学习一样，学习器容易受到孤立点(奇异样本)的影响，方法鲁棒性较低。

通过以上对各主动学习选择策略的介绍和相互对比可以看出，基于误差减少选择的主动学习方法，虽然学习精度高，但是其计算复杂，不适合在大数据集下贝叶斯网络的学习。而委员会投票选择的主动学习方法计算复杂度低，并且贝叶斯网络对未标注样本能够给出属于各种假设的后验概率，并据此可直接度量出样本类别的模糊性，为委员会投票选择样本提供了一种量化的度量标准，因此，QBC 主动学习方法更适于在大数据集下的贝叶斯网络分类器建模。

为了提高基于 QBC 主动学习的贝叶斯网络分类器的效率和性能，6.2.2 节介绍了作者对 QBC 主动学习方法的两种改进策略和基于这两种选择策略构建贝叶斯网络分类器的算法分析与实验。

6.2.2 基于主动学习的贝叶斯网络分类器学习算法

1. 基于投票熵与类条件后验最大熵相结合的 QBC 贝叶斯网络学习算法

(1) 算法原理

目前，已有的 QBC 主动学习算法中，度量委员会投票差异($D(x_i)$)有两种方法。其中

McCallum 和 Nigam 利用平均相对熵(Kullback-Leibler divergence to the mean, KL-dm)来度量委员会成员投票差异,计算公式为

$$D(\boldsymbol{x}_i)=\frac{1}{K}\sum_{m=1}^{K}D(p_m(Y\mid \boldsymbol{x}_i)\mid\mid p_{\text{avg}}(Y\mid \boldsymbol{x}_i)) \tag{6.12}$$

式中,$p_{\text{avg}}(Y|\boldsymbol{x}_i)$为所有委员会成员类条件概率的平均值,有

$$p_{\text{avg}}(Y\mid \boldsymbol{x}_i)=\left(\sum\nolimits_m p_m(Y\mid \boldsymbol{x}_i)\right)/K \tag{6.13}$$

$D(\cdot\mid\mid\cdot)$是两个条件概率分布的信息度量——相对熵,其计算公式为

$$D(p_1(Y)\mid\mid p_2(Y))=\sum_{j=1}^{|Y|}p_1(\boldsymbol{x}_j)\log\left(\frac{p_1(\boldsymbol{x}_j)}{p_2(\boldsymbol{x}_j)}\right) \tag{6.14}$$

相对熵越大,说明委员会成员的投票差异就越大。

Argamon-Engelson 使用了另一种度量方法——投票熵(Vote Entropy, VE),度量投票的不一致性。投票熵越大,说明委员会成员的投票差异越大。

$$D(\boldsymbol{x}_i)=-\frac{1}{\operatorname{logmin}(K,\mid Y\mid)}\sum_{Y}\frac{V(y,\boldsymbol{x}_i)}{K}\log\frac{V(y,\boldsymbol{x}_i)}{K} \tag{6.15}$$

式中,$V(y,\boldsymbol{x}_i)$为委员会成员对样本$\boldsymbol{x}_i$ 的类别 y 投票的数目。

表 6.1 和表 6.2 说明了两种度量方法对样本的评价结果。其中委员会成员 2 个,样本数据 3 个,两类分问题。表 6.1 中的数据为委员会成员对样本的类投票结果,表 6.2 是每个样本的类投票熵和平均相对熵的比较。从表 6.2 中可以看到,$\boldsymbol{x}_1$ 和$\boldsymbol{x}_3$ 的 KL-dm 值较接近,但是$\boldsymbol{x}_1$ 的 VE 值为 1,**KL-dm** 度量方法漏选了像x_1 这样委员会成员类投票不一致的样本,这样的样本正是 QBC 算法原理所要选择的例子。

表 6.1 委员会成员对样本的类投票结果

模型	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$
1	0.52(y_1)	0.72(y_2)	0.60(y_2)
2	0.58(y_2)	0.60(y_1)	0.70(y_2)

表 6.2 每个样本的类投票熵和相对熵

示例	VE	KL-dm
x_1	1	0.005(miss)
$\boldsymbol{x}_2$	1	0.052
$\boldsymbol{x}_3$	0	0.006

然而利用投票熵来度量委员会成员的投票不一致性,它虽然选择了投票不一致的例子,但是并未考虑成员对样本的类条件概率值 $p_m(y_j|\boldsymbol{x}_i)$,即没有考虑分类器成员对样本预测的类别可信度,故不能全面衡量样本所含信息量,这同样会导致漏选一些信息量丰富的样本数据。

下面给出定理 6.1,并证明基于投票熵的 QBC 算法漏选了投票一致且具有分类不确定性高的样本。

定理 6.1 设委员会成员数目 $k=2$,投票熵 $D(\boldsymbol{x})\in[0\quad 0.5\quad 1]$,其中 $D(\boldsymbol{x})=1$ 为投票不一致,$D(\boldsymbol{x})=0.5$ 为一个投票另一个弃权,$D(\boldsymbol{x})=0$ 为投票一致,委员会对样本的分类

不确定性为 $Uc_{\max}(\boldsymbol{x})$，则基于投票熵的 QBC 主动学习方法漏选的样本为 $\boldsymbol{x}_{\text{miss}} \in \{\boldsymbol{x} \mid D(\boldsymbol{x})=0, Uc_{\max}(\boldsymbol{x})>\alpha\}$，其中 α 为 $Uc_{\max}(\boldsymbol{x})$ 的阈值。

证明： 假设用函数 $Uc_m(\boldsymbol{x})$ 度量委员会成员 m 对样本 $\boldsymbol{x}_i$ 分类的不确定性，设样本 $\boldsymbol{x}$ 的类别属性值为 N 个，成员 m 对样本预测的类条件后验概率分布为 $p_m(Y \mid \boldsymbol{x})=(p_{m1}(y_1 \mid \boldsymbol{x}), p_{m2}(y_2 \mid \boldsymbol{x}), \cdots, p_{mN}(y_N \mid \boldsymbol{x}))$，很明显，如果概率分布为均匀分布时，成员 m 对样本的分类不确定性 $Uc_m(\boldsymbol{x})$ 达到最大，取成员中 $Uc_m(\boldsymbol{x})$ 的最大值记为 $Uc_{\max}(x)$ 的值。

因此，根据投票熵的定义式公式(6.15)，不失一般性，有以下 7 种情况：

① 如果 $p_{1i}>>p_{1\tilde{i}}, p_{2l}>>p_{2\tilde{l}}$，且 $i \neq l$，则 $D(\boldsymbol{x})=1, Uc_{\max}(\boldsymbol{x}) \leqslant \alpha$。

② 如果 $p_{1i} \approx \frac{1}{N}>p_{1\tilde{i}}, p_{2l}>>p_{2\tilde{l}}$，且 $i \neq l$，则 $D(\boldsymbol{x})=1, Uc_{\max}(\boldsymbol{x})>\alpha$。

③ 如果 $p_{1i} \approx \frac{1}{N}>p_{1\tilde{i}}, p_{2l} \approx \frac{1}{N}>p_{2\tilde{l}}$，且 $i \neq l$，则 $D(\boldsymbol{x})=1, Uc_{\max}(\boldsymbol{x})>\alpha$。

④ 如果 $p_{1i}>>p_{1\tilde{i}}, p_{2l}>>p_{2\tilde{l}}$，且 $i=l$，则 $D(\boldsymbol{x})=0, Uc_{\max}(\boldsymbol{x}) \leqslant \alpha$。

⑤ 如果 $p_{1i} \approx \frac{1}{N}>p_{1\tilde{i}}, p_{2l}>>p_{2\tilde{l}}$，且 $i=l$，则 $D(\boldsymbol{x})=0, Uc_{\max}(\boldsymbol{x})>\alpha$。

⑥ 如果 $p_{1i} \approx \frac{1}{N}>p_{1\tilde{i}}, p_{2l} \approx \frac{1}{N}>p_{2\tilde{l}}$，且 $i=l$，则 $D(\boldsymbol{x})=0, Uc_{\max}(\boldsymbol{x})>\alpha$。

⑦ 如果 $p_{1i}=\frac{1}{N}, p_{2l}>p_{2\tilde{l}}$，则 $D(\boldsymbol{x})=0.5, Uc_{\max}(\boldsymbol{x})>\alpha$。

其中，满足情况①、②、③和⑦的样本都被基于投票熵的 QBC 算法选择参与分类器的学习，而满足情况⑤和⑥的样本却被排除在选择之外。

由于 Lewis 和 Gale 在基于不确定性抽样选择的主动学习研究中指出，被分类器分类不确定的样本所含的信息量是丰富的，有利于分类器的构建，应当被选入训练集中。所以，满足情况⑤和⑥的样本因为具有较高的分类不确定性，应该被选择进行真实标注并加入到训练集中。

由此得出，基于投票熵的 QBC 主动学习方法漏选的样本为 $\boldsymbol{x}_{\text{miss}} \in \{\boldsymbol{x} \mid D(\boldsymbol{x})=0, Uc_{\max}(\boldsymbol{x})>\alpha\}$。

下面通过一个小的实验，举例证实了定理 6.1 的内容。

由于信息熵是信息论中描述事物不确定性的一个基本概念，因此，这里引入信息熵度量委员会成员对样本 $\boldsymbol{x}_i$ 分类的不确定性。其定义如下：

定义 6.1 [77] 设随机变量 x 是离散的，它取 $y_1, y_2, \cdots, y_k, \cdots$ 至多可列 $|Y|$ 个值，则

$$H(Y \mid x)=-\sum_{i=1}^{|Y|} p(y_i \mid x) \log (p(y_i \mid x)) \tag{6.16}$$

称为 x 的信息熵。

可以证明，随机变量(或随机向量)的信息熵最大的从充分必要条件是随机变量(或随机向量)为均匀分布。

在这里用信息熵 $H_m(Y \mid \boldsymbol{x})$ 度量了委员会成员 m 对样本分类的不确定性程度，取成员中的最大值记为样本的类条件后验最大熵 $H_{\max}(\boldsymbol{x})$ 的值，即

$$H_{\max}(\boldsymbol{x})=\max_{m=1}^{K} H_m(Y \mid \boldsymbol{x}) \tag{6.17}$$

式中，K 为委员会成员的数目。

假设有 2 个委员会成员，对 4 个未带类标注的样本$\boldsymbol{x}_1,\cdots,\boldsymbol{x}_4$进行预测，预测的类别有两个，即 $Y=\{y_1,y_2\}$。

利用投票熵度量委员会对样本类投票的不一致性 $D(\boldsymbol{x}_i)$，同时用类条件件后验最大熵 $H_{\max}(\boldsymbol{x})$度量成员对样本的分类不确定性。

表 6.3 中的数据为委员会成员对样本的类投票结果，表 6.4 是每个样本的类投票熵和类条件件后验最大熵的比较。从表 6.4 中可以看到，$\boldsymbol{x}_1$ 和$\boldsymbol{x}_4$ 的投票熵都为 0(投票一致)，但是前者的 $H_{\max}(\boldsymbol{x}_i)$值与$\boldsymbol{x}_2$ 的 $H_{\max}(\boldsymbol{x}_i)$值相当，即委员会中有对$\boldsymbol{x}_1$ 的分类不确定性程度相当高的成员。但是，基于投票熵的 QBC 算法漏选了这样的样本数据。

表 6.3 委员会成员对样本的类投票结果

模型	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$
1	0.55(y_1)	0.55(y_2)	0.72(y_1)	0.75(y_2)
2	0.60(y_1)	0.55(y_1)	0.85(y_2)	0.90(y_2)

表 6.4 每个样本的类投票熵和类条件件后验最大熵

示例	$D(\boldsymbol{x}_i)$	$H_{\max}(\boldsymbol{x}_i)$
$\boldsymbol{x}_1$	0.0(miss)	0.6881
$\boldsymbol{x}_2$	1.0	0.6881
$\boldsymbol{x}_3$	1.0	0.5930
$\boldsymbol{x}_4$	0.0	0.5623

该实验也说明了，基于投票熵的 QBC 算法漏选了投票一致且具有分类不确定性高的样本，这样的样本也具有丰富的信息量，有助于分类器性能的提高。

同时，QBC 同不确定抽样学习一样，学习器容易受到孤立点的影响，由于孤立点有较高的分类不确定性，因而易于被选择。如果学习器选择了孤立点，那么在接下来的参数调整过程中，当前分类器对它的分类误差也就越大，随着学习过程的进行，误差会继续传播下去，影响分类的性能。

而在主动学习过程中，如果选择当前假设对其标记比较确信的样本，则该样本标记不能提供较多有用信息，不能较多改变当前假设，但该操作可提高分类器鲁棒性。如果选择标记不能确定的样本，给定当前假设，该选择导致学习过程在两个或多个不同假设之间进行反复测试，可能会显著地改变当前假设，也可能会导致错误的更新，但该操作会提供更多有用信息，选择样本时应考虑两方面作用的平衡[70]。

基于投票熵的 QBC 算法，只度量分类器组对样本的投票不一致程度，选择当前假设分类最不确定的样本参与训练，会存在错误的更新，即受奇异样本的影响，因此应考虑适当选择确定性的样本，提高算法的鲁棒性。而该算法漏选的样本正是一些当前假设对其标记比较确信但却在微观层面上又相对不确定的样本，这些样本符合了确定性与不确定性相互平衡的要求。

因此，结合类条件件后验最大熵和投票熵，一种新的 QBC 算法由本书提出来，其中算法既度量委员会投票的不一致性，又对投票一致的样本进一步度量选择。选择投票不一致的样本，会提供更多有用信息，而选择相对确信中富信息量的样本，其既可以提高分类器鲁棒性

又提供了较多的分类信息，因此新算法选择样本时考虑了两方面作用的平衡。该算法中分类算法使用贝叶斯网络，算法伪代码如下：

Algorithm6.2 The QBC based on vote entropy and maximum entropy

```
QBC-VEME(
A,                                                      //分类算法
K,                                                      //委员会成员数目
L,                                                      //少量带有类别标注的样本集
U,                                                      //未带类别标注的候选样本集
ζ,                                                      //选择停止的条件
θ,                                                      //投票熵的阈值
α,                                                      //最大熵的阈值
)
1  While not ζ{
2   Train K classifiers {M_i: 1≤i≤K} from L based on A  //从 L 中学习 K 个分类器
3   For each x_i ∈ U
4       For h=1, …, K
5           Compute P(Y|x_i, M_h)         //基于 M_h 对x_i 进行分类预测，给出类标注 Y 的分布
6       End For
7       Compute D(x_i) with equation(6.15)//使用公式(6.15)计算 D(x_i)，度量委员会投票差异程度
8       If D(x_i) > θ
9           Select x_i from U to user for labeling         //从 U 中选择 x_i，交给用户去标注
10          Get the true label y_i of x_i and add (x_i, y_i) into L
                                                        //获得x_i 的真实类别，并将其加入到 L 中
11      Else
12          Compute H_max(x_i) with equation(6.16) and (6.17)
13          If H_max(x_i) > α
14                  Select x_i from U to user for labeling   //从 U 中选择x_i，交给用户去标注
15              Get the true label y_i of x_i and add (x_i, y_i) into L
                                                        //获得x_i 的真实类别，并将其加入到 L 中
16      End if
17   End for
18   Train classifier M from L                          //从 L 中训练分类器 M
19   Test M on test data and get its accuracy           // 测试分类器 M，得到它的分类精度
20 }
21 Return M
```

(2) 计算复杂度分析

下面对基于误差减少选择的主动学习方法、基于投票熵的 QBC 主动学习方法和基于投票熵与类条件后验最大熵相结合的 QBC 主动学习方法进行计算复杂度的比较。

假设一次内积计算的时间为 T（约为 $10\mu s$），训练贝叶斯网络分类器的时间为 T_c，未标注样本集包含 n 条数据样本，类别属性值为 n_y，那么基于误差损失选择的主动学习方法计算一次迭代的最长时间为 T_{ERS}，它等于 n 个未标注样本分别具有 n_y 个可能的类标签后加入训练集训练分类器的时间，计算分类损失的时间 $T_{\hat{R}_s}$，选取分类误差最小的未标注样本获得真实类标签后训练分类器的时间以及在测试集上计算的时间 T_{test} 的总和，即

$$T_{ERS} = n[n_y(T_c + T_{\hat{R}_s})] + T_c + T_{test} \tag{6.18}$$

因为,通常 n 趋向于很大,所以式(6.18)可写为

$$T_{\mathrm{ERS}} \approx [n(n_y(T_c + T_{\hat{R}_s})] \tag{6.19}$$

同时,$T_c \approx 3.410T$,$T_{\hat{R}_s} \approx 206.290T$,则

$$T_{\mathrm{ERS}} \approx 210 \cdot n \cdot n_y T \tag{6.20}$$

而基于投票熵的 QBC 主动学习方法计算一次迭代的最长时间为 T_{VE},它等于 k 个分类器训练的时间,计算投票差异的时间 T_{D},选择投票不一致的未标注样本加入训练集重新训练分类器的时间以及在测试集上计算的时间的总和,即

$$T_{\mathrm{VE}} = k \cdot T_c + T_{\mathrm{D}} + T_c + T_{\mathrm{test}} \tag{6.21}$$

式中,k 为委员会的成员数,其值为 2、3; $T_{\mathrm{D}} \approx 1.2T$;$T_{\mathrm{test}} \approx 54.68T$。

$$T_{\mathrm{VE}} \approx 70T \tag{6.22}$$

基于投票熵与类条件后验最大熵相结合的 QBC 主动学习方法计算一次迭代的最长时间为 $T_{\mathrm{VE}\&H_{\max}}$,它等于 k 个分类器训练的时间,计算投票差异的时间 T_{D},计算 $H_{\max}$ 的时间,选择 $H_{\max}$ 大于某一阈值的未标注样本加入训练集重新训练分类器的时间以及在测试集上计算的时间的总和,即

$$T_{\mathrm{VE}\&H_{\max}} = k \cdot T_c + T_{\mathrm{D}} + T_{H_{\max}} + T_c + T_{\mathrm{test}} \tag{6.23}$$

其中,取 $k=3$; $T_{\max} \approx 3T$。

$$T_{\mathrm{VE}\&H_{\max}} \approx 3T + T_{\mathrm{VE}} = 73T \tag{6.24}$$

因此,$T_{\mathrm{VE}} < T_{\mathrm{VE}\&H_{\max}} << T_{\mathrm{ERS}}$,所以,提出的新算法的计算复杂度要远小于基于误差减少选择的主动学习方法,实际应用是可行的。

另外,为了进一步再减少计算复杂度,还可以采取一些其他处理方法。例如,对未标注样本池进行抽样和聚类,以便只有类种子样本才考虑是否标注;也可以在每次增加新样本数据后,采用贝叶斯网络的增量学习方法,而不必从头开始训练贝叶斯网络。贝叶斯网络的增量学习将在 6.3 节中介绍。

(3) 实验结果和分析

实验使用 UCI 机器学习数据库中 tic-tac-toe 数据对基于投票熵的 QBC 和基于投票熵与类条件后验最大熵相结合的 QBC 及被动学习进行了结果比较,分类算法使用 TAN 分类器,委员会成员为 2 个。

tic-tac-toe 数据共有 958 条记录,有 2 个类别分类,其中 positive 类占 66.3%,negative 类占 34.7%。数据集被分成 3 个部分,其中随机抽取 59 条记录作为初始分类器的训练数据,299 条记录作为测试数据,剩下的 600 条记录被看作未带类别标记的样本。委员会成员通过随机抽取训练数据(Bagging)建立两个分类器来构建。投票熵的阈值 $\theta \geqslant 0.5$,$H_{\max}(\boldsymbol{x}_i)$ 的阈值 $\alpha > 0.6730$($\alpha = 0.6730$ 是类条件概率分布为 $p(x=a)=0.4$,$p(x=b)=0.6$),选择停止的条件 ζ 为分类准确性不小于 85%或者搜索完所有的未带类别标注的样本集。

图 6.5 所示为该组数据的结果图。从图中可以看到,基于投票熵的 QBC 方法(VE)选择停止后,总共选择了 163 条未带类别标注的样本,在选择了 138 条未带类别标注的样本数据时,分类预测准确率达到了其最高值 79%;基于投票熵与类条件后验最大熵相结合的 QBC 方法(VE& $H_{\max}$)选择停止后,总共选择了 183 个未带类别标注的样本,在选择了 176 条未带类别标注的样本数据时,分类预测准确率达到了其最高值 83%。被动学习使用了除测试数据以外的 659 条记录作为训练数据,分类准确率为 81%(表 6.5)。

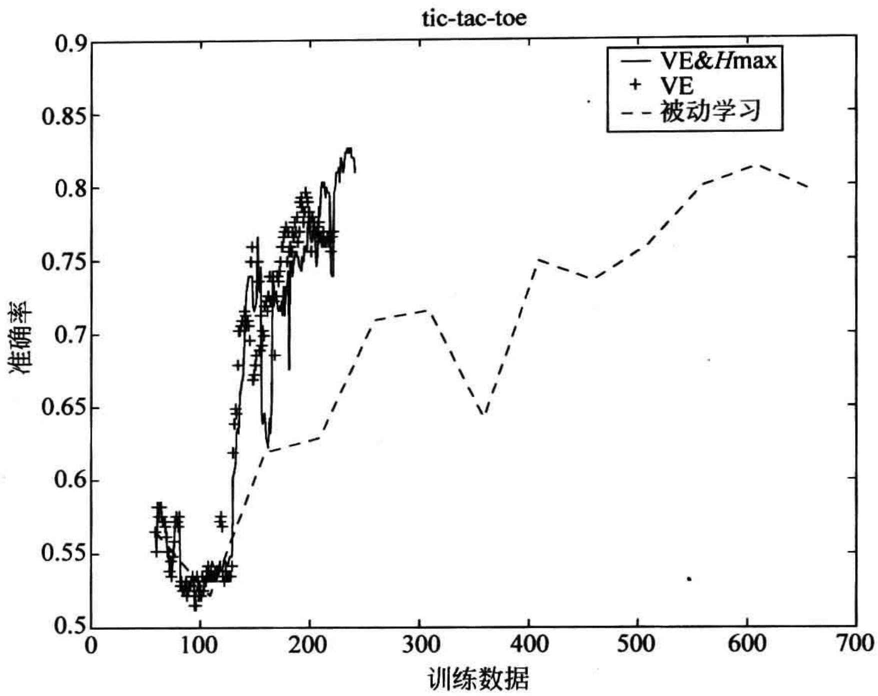


图 6.5 tic-tac-toe 数据 3 种学习方法分类准确率曲线

实验结果显示，基于投票熵与类条件后验最大熵相结合的 QBC 方法能够达到比基于投票熵的 QBC 更高的分类准确性，新加入的少量样本是含信息量丰富且相对确定的样本，有利于分类器性能的提高，减少了奇异样本的加入而增大的分类误差，并且算法使用了较少的训练数据来建立分类模型，减少了标注样本所付出的代价。

表 6.5 3 种学习方法的比较

学习方法	tic-tac-toe
VE	163(79%)
VE& H_{max}	183(83%)
被动学习	659(81%)

但是，信息熵是与计算系统中事件数的多少有关，它对两类分问题可以很好地度量样本分类不确定性，但对于多类分问题，并不利于对样本分类不确定性的度量。而相对熵是度量某概率分布密度偏离给定标准分布的程度，它有利于多类分概率分布密度与分类最不确定的概率分布相近程度的比较，相对熵越小表示两个分布的符合程度越好，成员对样本的分类越不确定。

例如，3 类分问题中，见表 6.6，$\boldsymbol{x}_3$ 的熵值要低于 $\boldsymbol{x}_1$，但 $\boldsymbol{x}_3$ 的分类不确定性要高于 $\boldsymbol{x}_1$。而 $\boldsymbol{x}_3$ 的相对熵值低于 $\boldsymbol{x}_1$，说明了它的分类不确定程度高于 $\boldsymbol{x}_1$。

表 6.6 样本的熵和相对熵

示例	$p_m(y_j \mid \boldsymbol{x}_i)$	熵	相对熵(与均匀分布时概率的比较)
$\boldsymbol{x}_1$	(0.6,0.2,0.2)	0.9503	0.1484
$\boldsymbol{x}_2$	(0.4,0.3,0.3)	1.0889	0.0098
$\boldsymbol{x}_3$	(0.55,0.45,0)	0.6881	0.005

2. 基于投票熵与最小相对熵相结合的 QBC 贝叶斯网络学习算法

(1) 算法原理

首先回顾相对熵(Kullback-Leibler divergence,KL-d)的一般定义。

定义 6.2 相对熵的离散形式是

$$\text{Min}: D(A \mid\mid B) = \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \tag{6.25}$$

式中,$A=(a_1,a_2,\cdots,a_n)$; $B=(b_1,b_2,\cdots,b_n)$; $a_i \geqslant 0$; $b_i \geqslant 0$,a_i 与 $b_i (i=1,2,\cdots,n)$ 是两个概率测度,且 $\sum_{i=1}^{n} a_i = 1$,$\sum_{i=1}^{n} b_i = 1$; $D(A \mid\mid B) = \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geqslant 0$ 为 A 相对于 B 的相对熵,$D(A \mid\mid B)$ 越小表示 A 与 B 的符合程度(一致性)越好。

在此考虑每个成员对样本分类的概率分布密度 $p_m(Y \mid \boldsymbol{x}_i)$,重新定义相对熵中的变量含义。

定义 6.3 在相对熵公式(6.25)中,让 A 代表 $p_m(Y \mid \boldsymbol{x}_i)$,$B$ 代表均匀分布时的概率分布 $P_{m_avg}(Y \mid \boldsymbol{x}_i)$,计算得到每个分类器对样本的相对熵,取其中最小值,得到该样本在分类器组中分类的最大不确定度,记为最小相对熵 KL-$\text{d}_{\min}(\boldsymbol{x}_i)$,其计算公式为

$$\text{KL-d}_{\min}(\boldsymbol{x}_i) = \underset{m=1}{\overset{K}{\text{Min}}}\left(\sum_{j=1}^{|Y|} P_m(y_j \mid \boldsymbol{x}_i) \log\left(\frac{P_m(y_j \mid \boldsymbol{x}_i)}{P_{m_avg}(Y \mid \boldsymbol{x}_i)}\right)\right) \tag{6.26}$$

式中,$P_{m_avg}(Y \mid \boldsymbol{x}_i) = \left(\sum_{j=1}^{|Y|} P_m(y_j \mid \boldsymbol{x}_i)\right) / l$,$l = \begin{cases} 2, \text{Max}(P_m(y_j \mid \boldsymbol{x}_i)) = 1 \\ (|Y| - \text{num_zero}), \text{Max}(P_m(y_j \mid \boldsymbol{x}_i)) \neq 1 \end{cases}$;

num_zero 为 $P_m(Y \mid \boldsymbol{x}_i)$ 为 0 的数目。

用相对熵度量样本类条件概率分布与分类最不确定时的概率分布的相近程度,计算所得相对熵越小,类条件概率分布与分类最不确定的概率分布越相近,样本分类不确定程度越高,最小相对熵体现了委员会成员对样本分类最大不确定程度。

现在使用表 6.7 所示的委员会对样本的投票结果(假设有 2 个委员会成员,对 3 个未带类标注的样本进行预测,预测的类别有 4 个),用投票熵度量委员会成员投票不一致性,用最小相对熵度量样本的分类不确定性,以便分析基于投票熵的 QBC 方法可能导致漏选的样本。

表 6.8 是每个样本的类投票熵和最小相对熵的比较。从表 6.8 中可以看到,$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_3$ 的投票熵都为 0,但是前者的 KL-$\text{d}_{\min}(\boldsymbol{x}_i)$ 值很接近于 0,即委员会中有对 $\boldsymbol{x}_1$ 的分类不确定性程度相当高的成员,应当被选入训练数据集中参与分类器的学习。

同基于投票熵与类条件后验最大熵相结合的 QBC 算法类似,基于投票熵和最小相对熵的 QBC 算法对投票一致的样本进一步选择,如果样本的 KL-$\text{d}_{\min}(\boldsymbol{x}_i)$ 满足某阈值条件,选择该样本加入训练数据集中。

表 6.7 委员会成员对样本的类投票结果

模型	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$
1	(0.4,0.3,0.2,0.1)(y_1)	(0.15,0.45,0.3,0.1)(y_2)	(0.1,0,0.8,0.1)(y_3)
2	(0.55,0.45,0,0)(y_1)	(0.75,0.1,0.1,0.05)(y_1)	(0.1,0,0.9,0)(y_3)

表 6.8 每个样本的类投票熵和最小相对熵

示例	VE	KL-d$_{min}$($\boldsymbol{x}_i$)
$\boldsymbol{x}_1$	0.0(miss)	0.005
$\boldsymbol{x}_2$	1.0	0.1509
$\boldsymbol{x}_3$	0.0	0.3681

基于投票熵和最小相对熵的 QBC 算法如下：

Algorithm6.3 The QBC based on vote entropy and minimum KL-d

```
QBC-VEKL(
A,                                                        //分类算法
K,                                                        //委员会成员数目
L,                                                        //少量带有类别标注的样本集
U,                                                        //未带类别标注的候选样本集
ζ,                                                        //选择停止的条件
θ,                                                        //投票熵的阈值
α,                                                        //最小相对熵的阈值
)
1 While not ζ{
2  Train K classifiers {M_i: 1≤i≤K} from L based on A  //从 L 中学习 K 个分类器
3  For each x_i ∈U
4      For h=1, …, K
5          Compute P(Y|x_i, M_h)          //基于 M_h 对x_i 进行分类预测，给出类标注 Y 的分布
6      End For
7      Compute D(x_i)with equation(6.15)   //使用公式(6.15)计算 D(x_i)，度量委员会投票差异程度
8      If D(x_i)> θ
9        Select x_i from U to user for labeling       //从 U 中选择x_i，交给用户去标注
10       Get the true label y_i of x_i and add (x_i, y_i) into L//获得x_i 的真实类别，并将其加入到 L 中
11     Else
12       Compute KL-d_min(x_i) with equation(6.26)
13       If KL-d_min(x_i) satisfy with α              //KL-d_min(x_i)满足 α 条件
14         Select x_i from U to user for labeling     //从 U 中选择x_i，交给用户去标注
15           Get the true label y_i of x_i and add (x_i, y_i) into L
                                                      //获得x_i 的真实类别，并将其加入到 L 中
16     End if
17  End for
18  Train classifier M from L                          //从 L 中训练分类器 M
19  Test M on test data and get its accuracy           // 测试分类器 M，得到它的分类精度
20 }
21 Return M
```

(2) 实验结果和分析

实验使用 UCI 机器学习数据库中 Nursery database 和 tic-tac-toe Endgame database 的数据对基于投票熵的 QBC 和基于投票熵与最小相对熵相结合的 QBC，及被动学习进行了结果比较。

实验从 Nursery database 中随机抽取了 4171 条数据记录，并把数据分成 3 个部分，其中 929 条记录作为初始分类器的训练数据，1001 条记录作为测试数据，剩下的 2241 条记录被看作未带类别标记的样本。整个数据集有 5 个类别分类，其类分布如表 6.9 所示。

tic-tac-toe 数据共有 958 条记录，有 2 个类别分类，其中 65%为类别是 positive，35%为 negative。把数据分成 3 个部分，其中随机抽取 59 条记录作为初始分类器的训练数据，299 条记录作为测试数据，剩下的 600 条记录被看作未带类别标记的样本。

表 6.9 Nursery database 的类分布

类	N	N/%
Not recom	1371	32.9
Recommend	1	0.02
Very recom	112	2.68
Priority	1391	33.35
Spec_prior	1296	31.05

委员会成员为 2 个，分类算法采用 TAN 贝叶斯网络分类器。

图 6.6 所示为 3 种学习方法对 Nursery database 数据的分类准确率比较，表 6.10 所示为结果对比。很明显，基于投票熵的 QBC 方法(VE)选择了 128 条未带类别标记的样本后准确率达到了 91%；VE&KL-d_{min}在选择了 264 条后，准确率为 93%；被动学习使用了所有的训练数据，准确率为 93%。

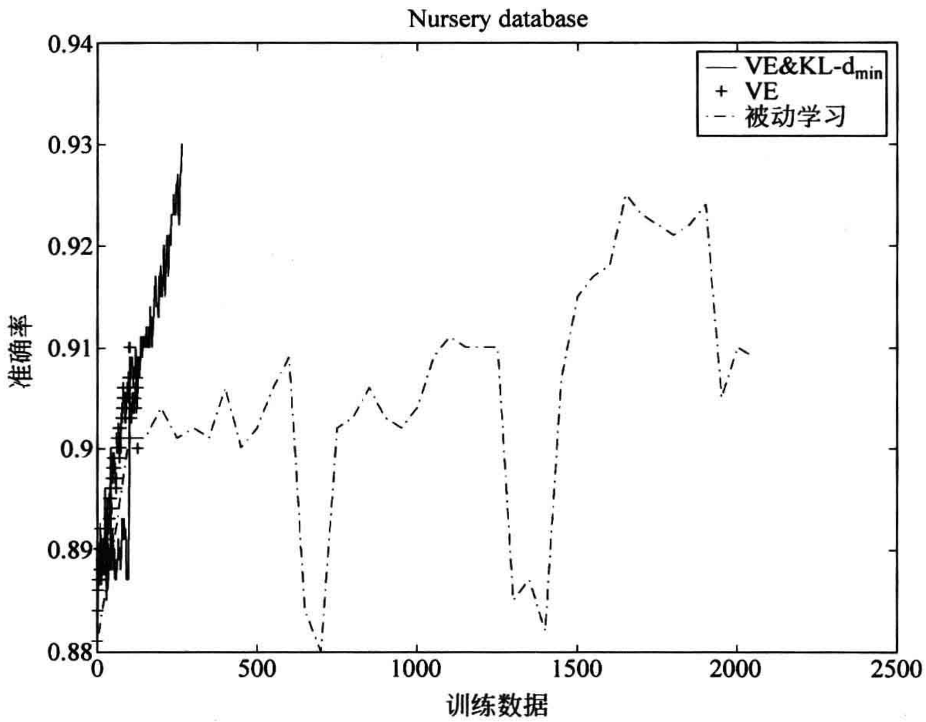


图 6.6 Nursery database 数据分类准确率曲线比较

(其中 $\theta \geq 0.5$、$\alpha > -0.0031$ 和 $\alpha < 0.05$、$\zeta \geq 95\%$)

表 6.10 Nursery database 3 种学习方法的比较

学习方法	Nursery database
VE	128(91%)
VE&KL-d_{min}	264(93%)
被动学习	2241(93%)

实验结果显示，对于多类分问题，基于投票熵与最小相对熵相结合的 QBC 方法能够达到比基于投票熵的 QBC 更高的分类准确性，新加入的少量样本是含信息量丰富且相对确定的样本，有利于分类器性能的提高，减少了奇异样本的加入而增大的分类误差，并且算法使用了较少的训练数据来建立分类模型，减少了标注样本所付出的代价。

图 6.7 所示为 4 种学习方法（VE、VE&KL-d_{min}、VE&H_{max}和被动学习）对 tic-tac-toe 数据的分类准确率比较，从图中可以看到，对于两类分问题，如果类分布比例较小，则 VE&KL-d_{min}方法分类表现不如 VE&H_{max}。

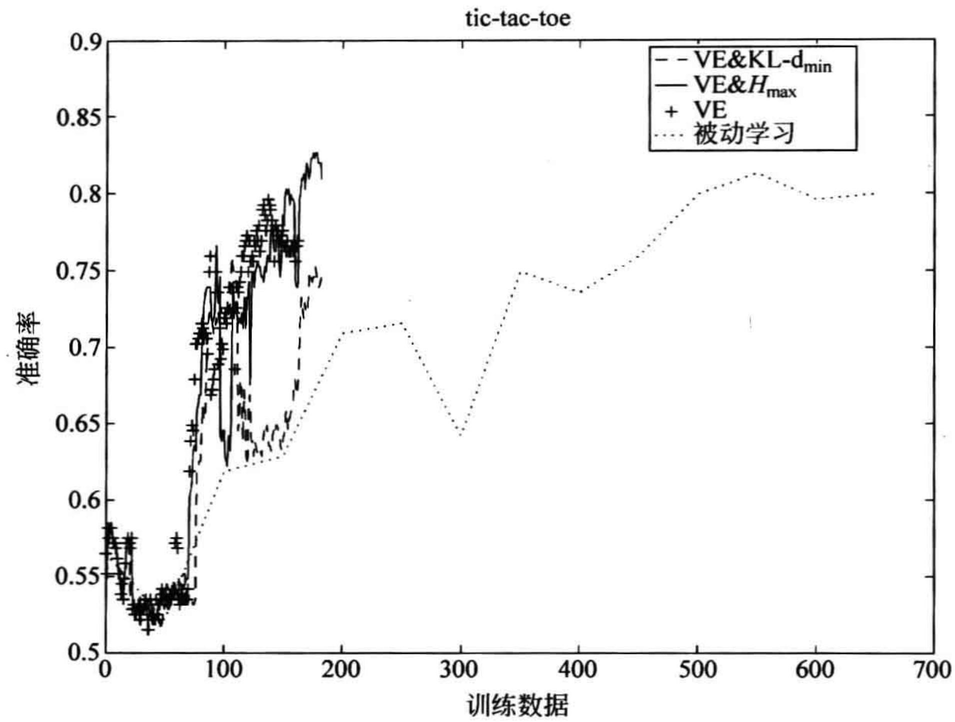


图 6.7 4 种学习方法（VE、VE&KL-d_{min}、VE&H_{max}和被动学习）对 tic-tac-toe 数据的分类准确率比较

通过 UCI 标准数据库的实验，结果表明，基于这两种新算法的主动贝叶斯网络分类器取得了与被动学习的贝叶斯分类器相同的分类精度，但是使用的训练数据远远少于被动学习。这无疑有效解决了大数据量训练样本获得困难情况下分类器建模困难和分类准确率不高的问题。

6.2.3 基于半监督主动学习的动态贝叶斯网络学习方法

动态贝叶斯网络是对具有随机过程性质的不确定问题进行建模和处理的一个有力工具，其用途十分广泛，如语音识别、表情识别、监测高速公路、客户信用状态的演化预测等。在针对分类任务的建模时，现有 DBNs 的学习算法都是基于大量带有类标记样本的监督学习，所建 DBNs 模型时间序列越长需要的学习样本也越多。然而，在大多数现实应用中，学习样本的类属性是需要人工标注的，对大量的学习样本作类别标记是一项枯燥而费时的工作。

虽然利用大量标注过的学习样本可以提高算法结果的准确度，但是却忽略了未标注样本的作用，而有效地利用未标注样本无疑将在一定程度上提高学习算法的性能。半监督学习(Semi-Supervised Learning，SSL)正是使用了既包括已标记类别样本又包括未标记类别

样本的学习方法，最经常用到的半监督学习算法是 EM 算法。近年来，研究者们对 DBNs 的半监督学习算法表现出与日俱增的研究兴趣，多种基于 EM 算法的 DBNs 学习方法已经被提出，如基于 SEM(Structural EM)算法的 DBN 结构学习、基于遗传算法和 EM 算法(EM-EA)相结合的 DBN 结构学习。虽然半监督学习算法只使用了少量的已标注样本，但其结果仍然不如基于大数据量的监督学习算法。例如，EM 算法是一种在有缺失值的情况下计算最大似然估计的迭代算法，它首先对未标注样本进行软分类，然后利用获得的完整样本集再进行模型的学习，显然未标注样本越多，相应的迭代次数也会越多，那么一旦在某些步骤加入了错误的样本分类信息，这类错误会随着迭代次数的增加而逐步的积累，最后必然会影响模型的准确性。

机器学习领域中的主动学习思想可以用来借鉴到基于半监督学习的 DBNs 中。主动学习器可以自主选择对学习过程最有价值的未标注样本来请求用户标记，达到以最少的样本尽最大可能优化当前学习器的目的。目前，多数主动学习算法在选择了所需标记的样本之后，不再对未标记数据作进一步利用。而半监督学习算法则可以在学习过程中利用所有未标记样本，将两种方法进行融合将是十分有价值的探索。

根据样本选择策略的不同，已有的主动学习算法可分为 3 类：一类是选择使得当前分类器对测试集分类误差最小的例子作为候选样本，如基于误差减少的抽样(Error Reduction Sampling，ERS) [84,85]；第二类是选择当前分类器最不确定的样本作为候选样本，如不确定性抽样方法(Uncertainty Sampling)；第三类是根据多个分类器对于样本类别预测差异程度来选择候选样本，如基于委员会投票选择方法(Query-By-Committee，QBC)。本章分别讨论了基于 QBC 的主动学习方法和 EM 算法相融合的 DBNs 算法以及基于 ERS 的主动学习方法和 EM 算法相融合的 DBNs 算法。总的研究结果表明，利用主动学习算法与 DBNs 的半监督学习算法相结合，能够提高 DBNs 学习的效率和性能，在少量的迭代次数后快速而准确地得到 DBNs 分类模型。

1. 基于误差减少和 EM 的动态贝叶斯网络学习算法

(1) 算法原理

EM 算法基于已标注和未标注样本学习得到动态贝叶斯网络分类器，由于在迭代过程中易于加入错误的样本分类信息，那么随着迭代次数的增加而误差逐步的积累，最后必然会影响模型的准确性。主动学习方法则可解决该问题，主动学习方法选择对于学习过程最为有用的样本，这些样本能够最大限度地减少分类器在未标注样本集上的分类误差，因而最大程度地减少了错误的分类信息的加入。

在监督学习中，用户所标记的样本是有冗余的，其中某些样本类别完全可通过分类器可靠地计算得到，半监督学习算法正是利用了该原则，来提高学习器性能。而半监督学习中引入主动学习方法，又能够使得分类器对未标注样本的软分类更加准确，从而最大程度地保证了分类信息的正确性。

动态贝叶斯网络分类器通过 EM 算法能够对未标记样本进行软分类，并且根据已标注样本集和软分类的样本重新学习 DBNs，通过得到的 DBNs 分类器可度量其在未标注样本池上的分类误差损失和，为主动学习算法中样本的选择提供了一种度量的依据。同时，由于主动学习可以自主挑选有用样本，减少学习所需要的实际标注样本数目，因而加快了学习

过程。

为了将基于误差减少的主动学习方法引入 DBNs 的 EM 半监督学习中，算法首先根据已标注样本集 L 和未标注样本集 U 学习动态贝叶斯分类器 H_L，在每次迭代过程中对任一未标注样本$\boldsymbol{x}_j$，考虑其所有可能的类标签 y_j，加$(\boldsymbol{x}_j, y_j)$到 L 中，学习另外一个分类器 H_{L^+}，根据公式(6.6)计算 H_L 和 H_{L^+} 在未标注样本池上的分类误差损失减少$\hat{R}_s$，选择使得$\hat{R}_s$ 最大的未标注样本来请求用户标注，并将标注后的样本加入到已有的训练集中。算法在新训练集上利用 EM 学习最终的分类器 H，如果该分类器的精度达到预定的值，则算法停止，输出最终分类器 H。该算法保证了学习的分类器的高准确性，并且减少人工标注样本的数目。其具体算法(DBNs-SSAL(ERS&EM))如下：

Algorithm6.4 The semi-supervised active learning based on error reduction and EM for DBNs

SSAL-EREM(
L, //少量带有类别标注的样本集
U, //未带类别标注的候选样本集
ζ, //选择停止的条件(这里是分类精度)
)

1 利用 DBNs-SEM 算法和 EM 算法，从 L 和 U 中学习 DBNs 分类器 H_L，并且在测试集上计算其分类准确性 ζ。
2 While not ζ{
3 对于 $\forall \boldsymbol{x}_j \in U$，考察其每一个可能的类标签 y_j，使得 $L^+ = L + (\boldsymbol{x}_j, y_j)$，$U^- = U - (\boldsymbol{x}_j)$，然后使用 DBNs-SEM 算法和 EM 算法从 L^+ 和 U^- 中学习分类器 H_{L^+}。
4 利用公式(6.6)计算$\hat{R}_s$。
5 对于 $\forall \boldsymbol{x}_j \in S = s_f(U) = \arg\max\limits_{S} \hat{R}_s$，从 U 中选择$\boldsymbol{x}_j$ 并且获得其真实的类标签 y_j^*，然后加入L，即 $L = L + (\boldsymbol{x}_j, y_j^*)$，$U = U - (\boldsymbol{x}_j)$。
6 使用 DBNs-SEM 算法 和 EM 算法从 L 和 U 中学习分类器 H，并且在测试集上计算其分类准确性 ζ。
7 }
8 返回 H

---

(2) 实验结果和分析

实验使用 MIT AI Lab 的 Bayes Net Toolbox for Matlab 中的一个 Water DBNs 的例子，随机抽取了部分数据进行实验。其中训练数据 2425 条，随机选取 25 条看作初始带类标签数据，其余 2400 条数据看作未带类标签数据，测试数据 1800 条。

实验对比了基于 DBNs-SSAL(ERS&EM)的学习算法与其他 3 种算法，即基于投票熵的 QBC 主动学习、基于半监督学习(Semi-supervised Learning)和基于监督学习(Supervised Learning)的动态贝叶斯网络分类器的分类准确性。实验结果如图 6.8 所示。从实验结果来看，基于 DBNs-SSAL(ERS&EM)算法在分类精度上好于主动学习(见图 6.9)、半监督学习，随着选择少量样本后其达到了与监督学习相同的分类精度。因此，分类性能上，超过了主动学习和半监督学习的 DBNs 分类器。但是，由于 DBNs-SSAL(ERS&EM)算法的计算复杂度相当高，运行了将近 167h 未得出最终结果，只得出前 8 个输出值，因此在计算效率上，不如主动学习和半监督学习。

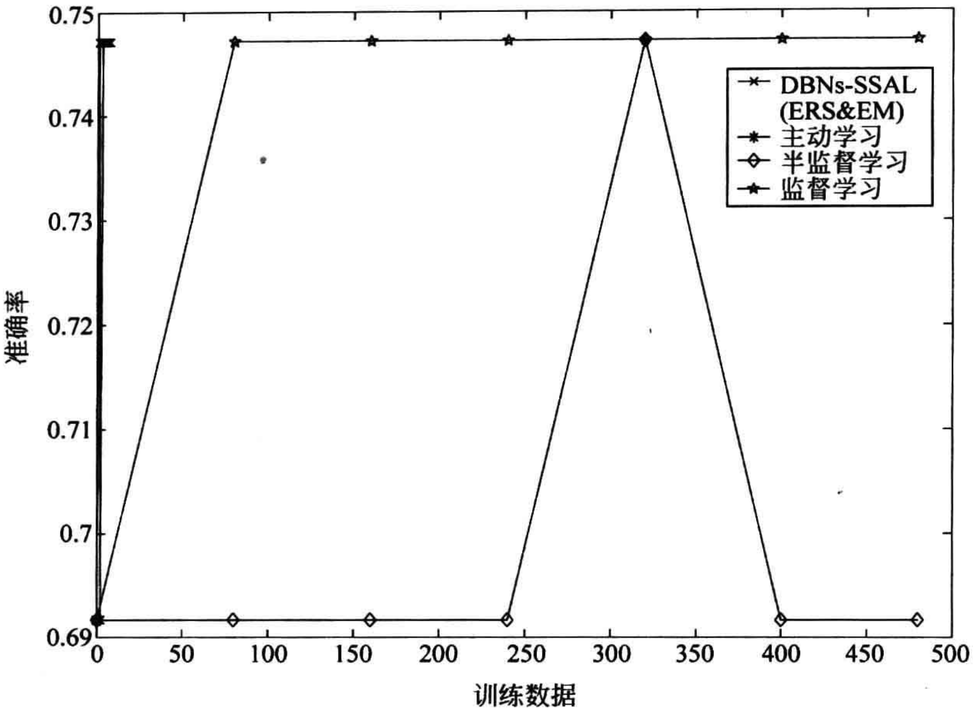


图 6.8 基于 DBNs-SSAL(ERS&EM)的学习算法与其他 3 种算法学习结果比较

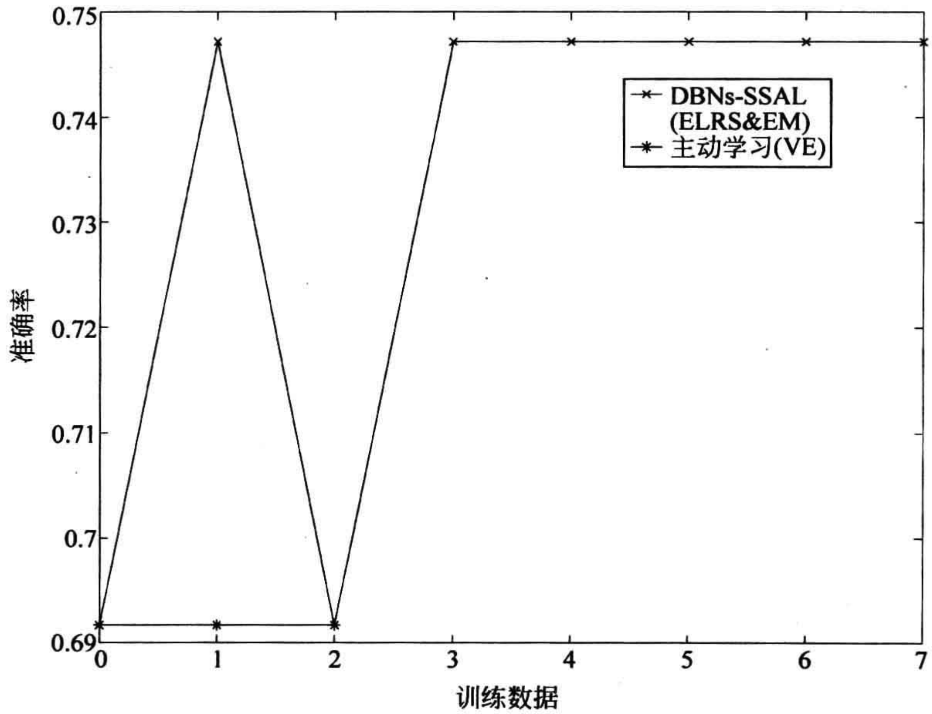


图 6.9 基于 DBNs-SSAL(ERS&EM)的学习算法与主动学习的比较

2. 基于 QBC 和 EM 的动态贝叶斯网络学习算法

(1) 算法原理

基于分类误差损失的主动学习方法，虽然其学习准确率高，但是选择了过多冗余样本，并且在每次选择样本之前必须搜索整个样本空间，计算每个样本的分类误差损失，才能确定哪些样本被选择，因此其学习时间长，计算复杂度高，对于利用大数据量未标注样本的半监

督学习算法并不可行。

然而,基于委员会投票选择的主动学习方法不需要检测整个样本空间,计算复杂度相对较低,学习速度快,但是,基于 QBC 的主动学习方法和不确定性抽样都易于选择奇异点样本——这些样本有较高的不确定性,奇异点样本的加入会使得分类器的分类误差加大,因此利用第 6.2.2 节提出的基于投票熵和最小相对熵相结合的 QBC 主动学习方法,以达到抑制由于不确定性样本的加入而产生的误差传播问题,提高 QBC 算法的鲁棒性。同时,利用该主动学习算法与 DBNs 的半监督学习算法结合,提高 DBNs 的学习效率和性能。

基于 QBC 的主动学习方法首先根据已有的类标签数据建立两个或多个分类器,组成"委员会",利用这个委员会对预测样本进行分类投票,然后选择投票最不一致的样本作为候选样本。这种方法能够选择对于学习过程最为有用的样本,从这些样本中获得丰富的分类信息,加快学习过程,因此它能够使用很少的已标记样本达到给定的分类精度,减少了人工标注的代价。但是,由于奇异点样本有较高的分类不确定性,因此委员会对其投票差异较大,学习器易于选择这样的样本加入训练集,使得分类误差加大。

然而,基于投票熵和最小相对熵相结合的 QBC 算法,其在选择投票不一致样本的同时,再选择一些分类相对确定并含有丰富分类信息的样本,以便弥补分类误差的增大,从而提高分类器的准确性。

基于 QBC 和 EM 的动态贝叶斯网络学习算法,首先根据已标注样本集 L 学习 K 个动态贝叶斯分类器,组成投票委员会,在每次迭代过程中对任一未标注样本进行分类投票,如果投票不一致,选择该样本进行标注并加入 L 中,利用 DBNs-SEM 算法和 EM 算法,从 L^+ 和 U^- 中学习分类器 H;如果投票一致,根据公式(6.26)考察其分类信息的含量,如果满足某一阈值,则重复和投票不一致时相同的操作步骤。重复迭代直到满足停止条件。其具体算法(DBNs-SSAL(VE-KLdmin&EM)如下:

Algorithm6.5 The semi-supervised active learning based on QBC and EM for DBNs

SSAL-QBCEM(
L, //少量带有类别标注的样本集
U, //未带类别标注的候选样本集
ζ, //选择停止的条件(这里是分类精度)
θ, //投票熵的阈值
α, //最小相对熵的阈值
)
1 While not ζ{
2 利用 DBNs-SEM 算法和 EM 算法,在 L 上学习 K 个 DBNs 分类器$\{h_1,\cdots,h_k\}$;
3 对于 $\forall \boldsymbol{x}_j \in U$,用$\{h_1,\cdots,h_k\}$软分类。
4 利用公式(6.15)计算 $D(\boldsymbol{x}_j)$。
5 如果 $D(\boldsymbol{x}_j) > \theta$。
6 从 U 中选择$\boldsymbol{x}_j$,获得其真实类别 y_j^*,并加入到 L 中,使 $L=L+(\boldsymbol{x}_j, y_j^*)$,$U=U-(\boldsymbol{x}_j)$,转到第 9 行。
7 如果 $D(X_j) \leqslant \theta$,利用公式(6.26)计算 KL-$\mathrm{d}_{\min}(\boldsymbol{x}_j)$。
8 如果 KL-$\mathrm{d}_{\min}(\boldsymbol{x}_j)$满足 α,转到第 6 行。
9 利用 DBNs-SEM 算法和 EM 算法,在 L 和 U 上学习 DBNs 分类器 H
10 }
11 返回 H

---

(2) 实验结果和分析

实验使用 MIT AI Lab 的 Bayes Net Toolbox for Matlab 中的一个 Water DBNs 的例子，随机抽取了部分数据进行实验。其中训练数据 2425 条，随机选取 25 条看作初始带类标签数据，其余 2400 条数据看作未带类标签数据，测试数据 1800 条。

实验对比了基于 DBNs-SSAL(VE-KLd_{min} &EM) 的学习算法与其他 3 种算法，即基于投票熵的 QBC 主动学习、基于半监督学习和基于监督学习的动态贝叶斯网络分类器的分类准确性。实验结果如图 6.10 所示。基于 DBNs-SSAL(VE-KLd_{min} &EM) 算法运行了 47.2 分钟，从实验结果来看，基于 DBNs-SSAL(VE-KLd_{min} &EM) 算法前期表现与主动学习相同（放大图见图 6.11），但是随着再选择少量样本后其达到了与监督学习相同的分类精度。因此，无论是效率和性能上，超过了其他 3 种版本的 DBNs 分类器。

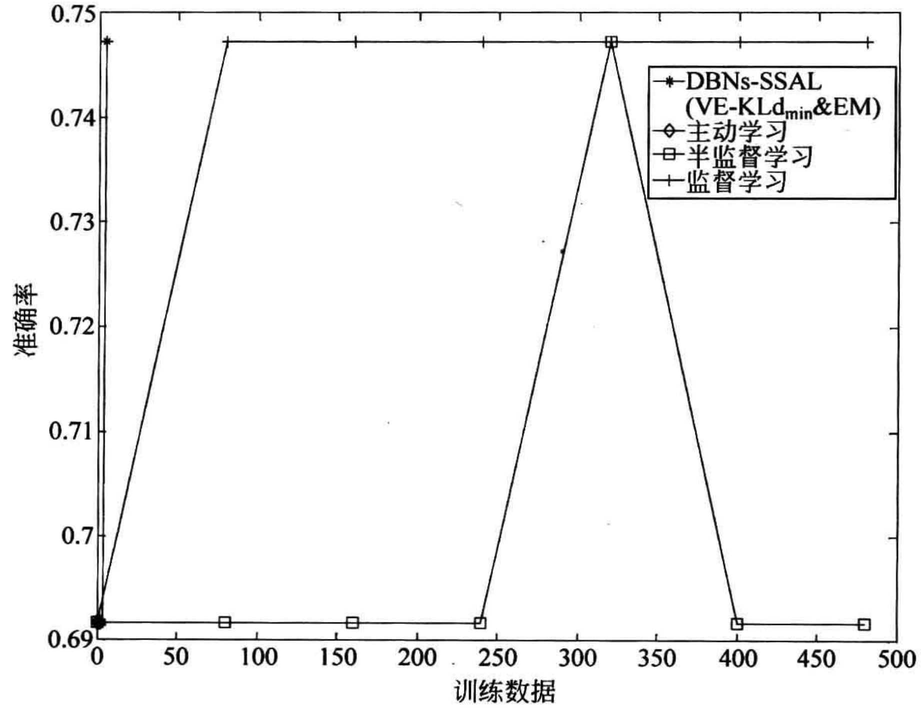


图 6.10 基于 DBNs-SSAL(VE-KLd_{min} &EM) 的学习算法与其他 3 种算法学习的学习结果比较

下面的实验比较了基于 DBNs-SSAL(VE-KLd_{min} &EM) 算法和基于 DBNs-SSAL (VE&EM) 算法。实验结果如图 6.12 所示，从实验结果来看，基于该数据集的 VE&EM 算法表现与基于投票熵的 QBC 主动学习的表现相同，VE-KLd_{min} &EM 算法比 VE&EM 算法所建分类器的性能更好，可以达到与监督学习相同的分类精度。因此，基于 DBNs-SSAL (VE-KLd_{min} &EM) 算法能够在少量的迭代次数后快速而准确地得到 DBNs 分类模型。

从计算复杂度上来分析，基于 DBNs-SSAL(VE-KLd_{min} &EM) 算法要比基于 DBNs-SSAL(ERS&EM) 算法复杂度低，实现容易，并且其两者分类表现相当。

6.2.4 基于主动学习的贝叶斯网络结构学习

通常学习贝叶斯网络结构需要大量的训练样本数据，训练样本数据需要进行人工标注并仔细地挑选，因此十分耗费人力和物力。2001 年由 Simong Tong 和 Daphne Koller[58] 提出的基于主动学习的贝叶斯网络学习方法能够有效地解决这一问题。其利用主动学习机制

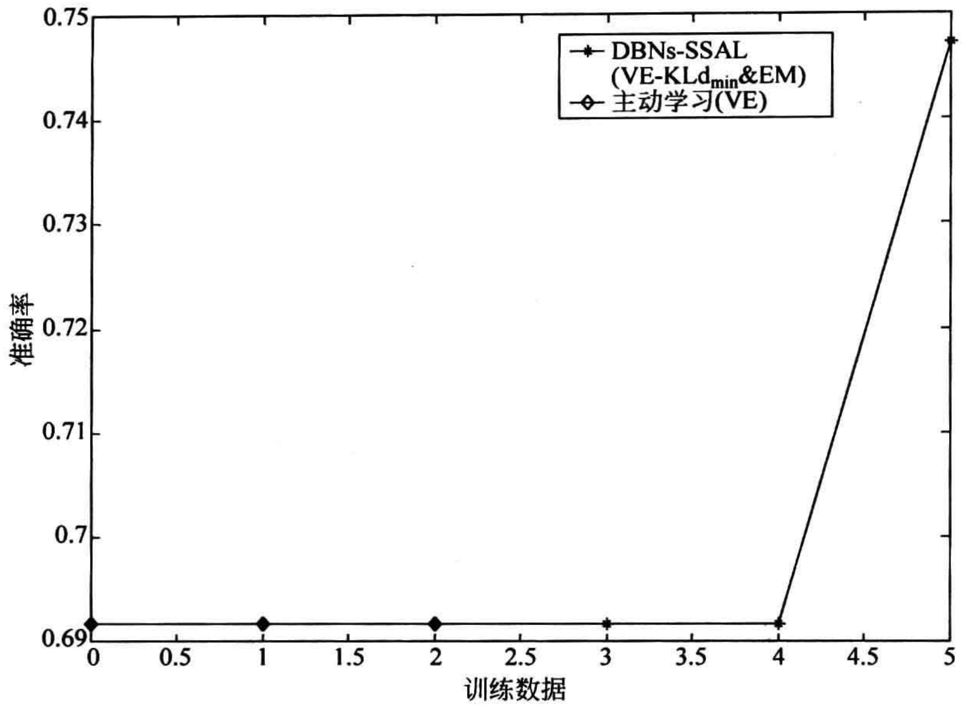


图 6.11 基于 DBNs-SSAL(VE-KLd_{min}&EM)算法与主动学习的学习结果比较

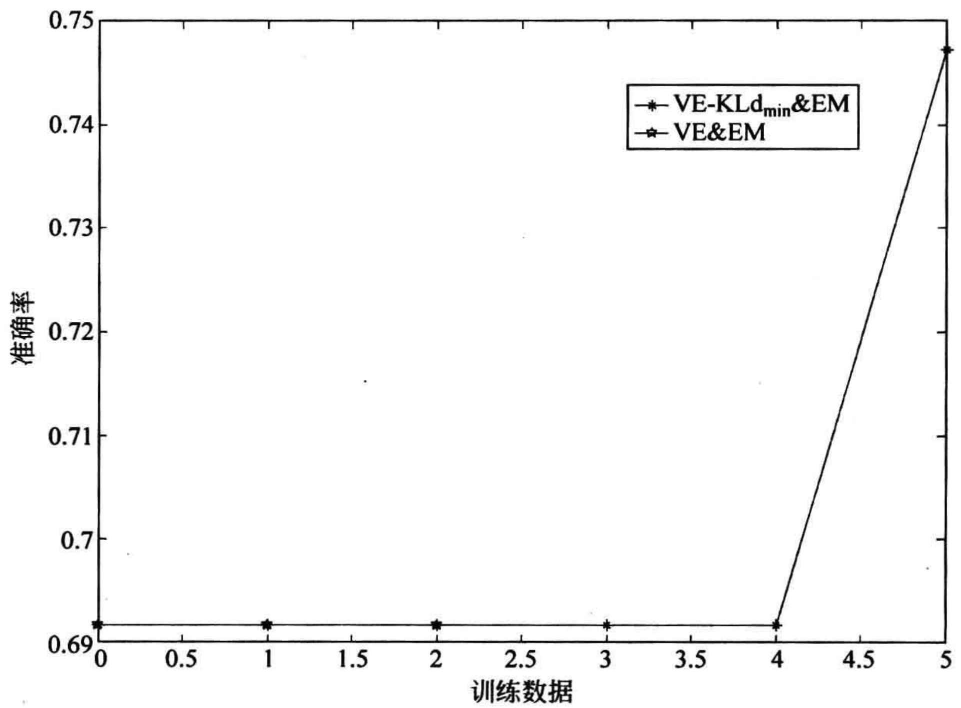


图 6.12 基于 DBNs-SSAL(VE-KLd_{min}&EM)算法与基于 DBNs-SSAL(VE&EM)算法的学习结果比较

决定最有信息价值的学习过程来发现潜在的最佳模型结构。主动学习器在结构学习过程中通过设置一些变量的值来介入学习过程。Tong 和 Koller 描述的算法主要是基于当前的模型结构,来主动地选择下一步需要的训练样本,而不是被动地按样本顺序接受数据来学习结构。他们的实验结果表明主动学习能够大幅度减少所需训练数据量,并且同被动学习相比,基于相同数据量其能够学习更为准确的贝叶斯网络结构。

下面介绍该算法的基本原理和实验效果。

1. 贝叶斯网络的介入查询学习

假定一个贝叶斯网络(G,θ_G)，其用一组随机变量$\chi=\{X_1,\cdots,X_n\}$来表示结点变量，用$\mathrm{Val}[X_i]$表示变量的取值范围。由于贝叶斯网络是一个概率模型，它可以回答形如$P(X_i|\mathrm{Pa}(X_i)=u_i)$的查询问题，其中$u_i$表示$\mathrm{Pa}(X_i)$的一组取值。同时，贝叶斯网络也可以看作是一个因果模型，因此其可以被用来回答介入查询，这使得在介入模型后，即使某些变量被迫等于某个特定值时，也能够得到相关的概率分布。

介入动作很可能导致删除变量的某个边，而破坏了模型的结构。也就是说，在介入动作以后，模型中结点X没有了父结点，然而真实模型中，$X=x$通过证据推理可能给出关于其父结点的一些信息，但是在人介入学习过程中，$X=x$的事实并没给出任何关于其父结点取值的任何信息。比如，对于一个汽车故障诊断模型，如果观察到电池没有充电，那么可能会得出发电机传输带出问题了。但是如果故意抽干电池，那么电池是空的事实不会提供任何关于发电机传输带的信息。因此，如果设置$X:=x$，那么结果模型的分布中，由于破坏了结构G(即删除了到X的边)，需要设置这些结点的CPDs，以便使$X=x$时的概率是1。

文献[58]算法的目标是从数据中学习结构G^*。很显然，给定足够的数据，是能够重新构造出模型的概率分布P^*的。然而，一般来说，P^*并不是唯一地决定模型结构G。例如，如果网络结构G^*有形式$X\longrightarrow Y$，那么对于相同的P^*也能够等价得出$Y\longrightarrow X$。对于给定的来自P^*的样本，希望最好能够得到G的马尔可夫等价类[299]：一组能够准确地推导出相同独立假设的网络结构集合。在马尔可夫等价类中，网络的框架(连接(X,Y)的边集合)是固定的，但是对于一些边，有的边的方向是固定的，而其他边的方向是任意的。如果试图得到X和Y之间的边的方向，那么在X处介入，改变其取值，看到关于Y的分布并没有变化，但是在Y处介入时，关于X的分布变化了，就此会得出边的方向是$Y\longrightarrow X$，因为Y的变化而导致了X的变化。

因此，文献[58]的算法是利用介入机制来学习变量之间的关系。其通过主动学习器选择某些变量进行介入，控制这些变量的取值，并通过查询，获得在该介入情形下其他变量的取值结果，通过获取的数据样本进一步学习网络结构，算法不断重复这个过程，直到收敛。在这里，算法用Q表示查询变量的子集，主动学习器能够为Q选择特别的取值q。请求$Q:=q$叫做查询，查询结果被称为响应，其是一个所有非查询变量随机抽样的样本x。换句话说，x是一个查询的结果，通过设定Q取q值来介入时，x是从上面描述的被破坏的模型中抽样而来。

那么介入查询获得新的数据样本(x,q)以后，算法是如何更新网络结构和参数分布的呢？

首先算法假设CPD参数是多项式形式和共轭Dirichlet分布。

给定一个在χ上完整的抽样样本d，使用贝叶斯规则得到在结构G上的后验分布是$P(d|G)P(G)$。$P(d|G)$可以被表达成在G上的所有可能参数值的积分，即

$$P(d\mid G)=\int P(d\mid G,\theta_G)p(\theta_G\mid G)\mathrm{d}\theta_G \tag{6.27}$$

如果d不是一个完整的抽样数据，当进行了一个介入查询$Q:=q$，并且响应结果是获得了数据x，算法通过恒等式(6.28)来更新概率分布$P(G,\theta_G)$，即

$$P(G,\theta_G \mid Q:=q,x)=p(\theta_G \mid Q:=q,x,G)\cdot P(G \mid Q:=q,x) \tag{6.28}$$

从式(6.28)可以看出,获得新的贝叶斯网络需要更新参数密度和更新结构的分布。

对于参数的更新,由于介入动作已经使 Q 等于某个固定值,所以只需要利用新的数据样本更新非查询变量的参数。对于更新结构分布,该算法使用公式(6.29)来计算 $P(G|Q:=q,x)$,其来自 Cooper 和 Yoo 的文献[300]中的定理,即

$$\begin{aligned}P(G \mid Q:=q,x) &= \frac{P(x \mid Q:=q,G)P(G)}{P(x \mid Q:=q)} \\ &= \frac{P(G)}{P(x \mid Q:=q)}\prod_{i:X_i\notin Q}\mathrm{Score}(X_i,U_i^G \mid x,q)\end{aligned} \tag{6.29}$$

其中,

$$\mathrm{Score}(X_i,U \mid d)=\int P(x_i \mid u,\theta_{X_i|u})p(\theta_{X_i|u})\mathrm{d}\theta_{X_i|u}=P(x_i \mid u) \tag{6.30}$$

x_i 和 u 是 X_i 和 U 在数据样本 d 中的值。

2. 主动学习思想

文献[58]的算法的主要目标不只是基于介入数据更新分布,其重点是利用主动学习方法主动地选择数据,这些数据可以使学习器学习到最佳的网络结构。一个近视(Myopic)的主动学习器 l 是一个函数,其可以基于当前结构和参数的分布来选择一个查询 $Q:=q$,进而获得响应数据 x,以便更新结构和参数的分布。主动学习器不断重复这个过程,直到算法收敛。之前描述了该算法如何更新结构和参数的分布,现在介绍一下该算法是如何构造这个主动学习器,以达到基于当前的分布 P 来选择下一个查询。

正如 Tong 和 Koller(76)文中的研究工作一样,构造主动学习器的关键步骤是定义一个评价函数,用它来评价网络结构和参数分布得好坏。通过使用这个评价函数来评价究竟哪些数据样本可能提高网络的分布质量,进而选择下一个查询来进行结构学习。

给定一个网络结构和参数的分布 $P(G,\theta_G)$,该算法使用损失函数 Loss(P)来度量结构和参数分布的质量。给定一个查询 $Q:=q$,算法定义了查询的期望后验损失为

$$\mathrm{ExPLoss}(P(G,\theta_G) \mid Q:=q)=E_{x\sim P(x|Q:=q)}\mathrm{Loss}(P(G,\theta_G \mid Q:=q,x)) \tag{6.31}$$

基于这个定义,一个简单的算法是:对于每一个候选查询 $Q:=q$,评估其期望后验损失,然后选择度量值最低的那个查询。然而,期望后验损失的计算复杂度相当高。因为算法需要在结构集合上获得一个分布,网络结构的数目是结点数目的超指数,并且给定一个查询,为了计算期望后验损失,算法不得不在整个结构集合上对可能的每一个指数级响应进行计算。为了确保学习算法的收敛,必须合理定义一个损失函数。

由于该算法的主要目标是学习正确的网络结构,因此,算法只对图模型的形式和每一个有向边感兴趣。对于两个结点 X_i 和 X_j,有 3 种可能的边的关系,即 $X_i\to X_j$、$X_i\leftarrow X_j$、X_i X_j。由于,利用 $P(G,\theta_G)$可以导出这 3 种可能的边的关系的一个分布,因此,通过对推导出的这个分布求取熵值,可以评价两个结点之间的关系的确定程度,即

$$\begin{aligned}H(X_i\leftrightarrow X_j)=&-P(X_i\to X_j)\log P(X_i\to X_j)\\&-P(X_i\leftarrow X_j)\log P(X_i\leftarrow X_j)\\&-P(X_iX_j)\log P(X_iX_j)\end{aligned} \tag{6.32}$$

这个熵值越大,关于结点 X_i 和 X_j 之间的关系越不确定。式(6.32)构成了主动学习器

边熵损失函数的基础，有

$$\mathrm{Loss}(P(G,\theta_G)) = \sum_{i,j} H(X_i \leftrightarrow X_j) \tag{6.33}$$

通过对分布 $P(G,\theta_G)$ 定义了损失函数后，Tong 和 Koller 提出一个有效的算法来计算给定的查询 $Q:=q$ 的期望后验损失。注意到 $P(G,\theta_G)$ 是当前的分布，其是由目前所获得的所有数据决定的。在算法的初始，这是一个先验分布，随着算法不断获得更多的数据，主动学习器使用如前所描述的方法来更新 $P(G,\theta_G)$，然后应用这个更新的 $P(G,\theta_G)$ 再作为新的先验分布，不断重复这个过程。

3. 结点顺序分析

为了使算法可执行，Tong 和 Koller 又引入结点顺序的一个分布，即网络结构遵循某个完全的结点顺序 $\prec$。这个问题分为了两种情况：

(1) 一个固定的结点顺序

让 $\prec$ 是 χ 的一个完全顺序。只关注那些满足 $\prec$ 的网络结构，即如果存在边 $X \to Y$，那么 $X \prec Y$。假设在每一次查询迭代之前，每一个结点 X_i 有 m 个最可能的固定父结点集 W_i。结点固定顺序的限制，主要作用在于：一个结点的父结点选择独立于其他结点父结点的选择(301 和 302)。文献[58]给出了下边两个重要的结论，即给出了查询响应 x 的概率分布和边的概率分布的有效计算表达形式。

定理 6.2 给定一个查询 $Q:=q$，对于这个查询响应 x 的概率为

$$\begin{aligned} P(x \mid Q:=q, \prec) &= \sum_{G \in \prec} \prod_i P(\mathrm{Pa}(X_i)=U_i^G) \prod_{j:X_j \notin Q} \mathrm{Score}(X_j, U_j^G \mid x, q) \\ &= \lambda_Q \prod_{i:X_i \notin Q} \sum_{U \in u_{i,\prec}} P(\mathrm{Pa}(X_i)=U)\mathrm{Score}(X_i, U \mid x, q) \end{aligned} \tag{6.34}$$

其中，$\lambda_Q = \prod_{i:X_i \in Q} \sum_{U \in u_{i,\prec}} P(\mathrm{Pa}(X_i)=U)$。

定理 6.3 给定一个查询 $Q:=q$ 和其完整的响应 x，边 $X_j \to X_i$ 的概率为

$$P(X_j \to X_i \mid Q:=q, x, \prec) = \frac{\sum_{U \in u_{i,\prec}, X_j \in U} P(\mathrm{Pa}(X_i)=U)\mathrm{Score}(X_i, U \mid x, q)}{\sum_{U \in u_{i,\prec}} P(\mathrm{Pa}(X_i)=U)\mathrm{Score}(X_i, U \mid x, q)} \tag{6.35}$$

其中，定义如果 $X_i \in Q$，则 $\mathrm{Score}(X_i, U \mid x, q)=1$。

现在考虑给定 $\prec$ 的期望后验损失：

$$\mathrm{ExPLoss}_{\prec}(P(G,\theta_G) \mid Q:=q) = E_{x \sim P(X \mid Q:=q,\prec)} \sum_{i,j} H(X_i \leftrightarrow X_j \mid Q:=q, x, \prec) \tag{6.36}$$

使用定理 6.3 计算 $H(X_i \leftrightarrow X_j \mid Q:=q,x,\prec)$，并且从该定理可以注意到，$H(X_i \leftrightarrow X_j \mid Q:=q,x,\prec)$ 仅仅依赖于 q 和 x 给 X_i、X_j、W_i 和 W_j 的值。使用这个事实那么应用定理 6.2，可以重新把期望后验损失公式(6.36)写成

$$\begin{aligned} \mathrm{ExPLoss}_{\prec}(P(G,\theta_G) \mid Q:=q) &= E_{x \sim P(X \mid Q:=q,\prec)} \sum_{i,j} H(X_i \leftrightarrow X_j \mid x_i, x_j, W_i, W_j, \prec) \\ &= \sum_{i,j} \sum_x P(x \mid Q:=q, \prec) H(X_i \leftrightarrow X_j \mid x_i, x_j, W_i, W_j, \prec) \end{aligned}$$

$$= \sum_{i,j} \sum_{x} H(X_i \leftrightarrow X_j \mid x_i, x_j, W_i, W_j, \prec) \times \lambda_Q \prod_{k: X_k \notin Q} \sum_{U \in u_{k,\prec}} P(\mathrm{Pa}(X_k))$$

$$= U\ \mathrm{Score}(X_k, U \mid x_k, W_k)$$

$$= \lambda_Q \sum_{i,j} \sum_{x} \psi(x_i, x_j, W_i, W_j) \prod_{k: X_k \notin Q} \phi(x_k, W_k) \tag{6.37}$$

其中，

$$\psi(x_i, x_j, W_i, W_j) = H(X_i \leftrightarrow X_j \mid x_i, x_j, W_i, W_j, \prec) \tag{6.38}$$

$$\phi(x_k, W_k) = \sum_{U \in u_{k,\prec}} P(\mathrm{Pa}(X_k)) = U\ \mathrm{Score}(X_k, U \mid x_k, W_k) \tag{6.39}$$

可以使用标准的图模型推理过程[303]有效评估公式(6.37)。

(2) 不限制结点的顺序

先前对于给定的结点顺序，关于一个查询获得了一个闭合形式表达的期望后验损失计算公式。现在，通过忽略对结点顺序的限制，来考虑一般化问题，那么期望后验损失能被重新写回下面的表达形式，即

$$\begin{aligned} &\mathrm{ExPLoss}(P(G,\theta_G) \mid Q:=q) \\ &= E_{x\sim P(X|Q:=q)} \mathrm{Loss}(P(G,\theta_G \mid Q:=q,x)) \\ &= E_{\prec}\ E_{x\sim P(X|Q:=q,\prec)} \mathrm{Loss}(P(G,\theta_G \mid Q:=q,x)) \\ &= E_{\prec}\ E_{x\sim P(X|Q:=q,\prec)} \sum_{i,j} H(X_i \leftrightarrow X_j \mid Q:=q,x) \end{aligned} \tag{6.40}$$

顺序的期望值 $E_{\prec}$ 可以用从当前结构和参数的分布，抽样可能的顺序来近似。正如 Friedman 和 Koller[302]中指出的，使用 MCMC 技术可以有效计算顺序的抽样。

4. 算法的描述

现在对 Tong 和 Koller 的基于主动学习的贝叶斯网络结构学习算法做一总结。该算法首先从模型和参数的当前分布抽样一套顺序集合，然后使用这套顺序集合计算每一个可能查询的熵 $H(X_i \leftrightarrow X_j \mid Q:=q,x)$。接着，对每一个顺序，使用标准的贝叶斯网络推理算法计算 $E_{x\sim P(X|Q:=q,\prec)} \sum_{i,j} H(X_i \leftrightarrow X_j \mid Q:=q,x)$，以获得在每一个可能查询上的因子 $h_{\prec}(Q)$。接着，针对每一个可能的查询，所有不同顺序的查询因子求平均值。

例如，如果有3个顺序，分别获得了3个因子 $h_{\prec_1}(Q)$、$h_{\prec_2}(Q)$、$h_{\prec_3}(Q)$，接着需要产生一个新的因子 $h(Q)$，在这个因子中，每个 $h(q)$ 实体是3个原始因子中 q 实体的平均。平均这些因子的过程就是计算所有顺序的期望。最后的结果是一个查询因子 $h(Q)$，对于查询变量 Q 的每一个可能的值 q，该因子给出了相应的期望后验损失值，最后选择期望后验损失值最低的查询进行响应。

贝叶斯网络结构的主动学习算法描述如下：

Algorithm6.6 The active learning for BN structure

Active Learning($P(G,\theta_G)$)

1 Sample orderings using MCMC //使用 MCMC 抽样一组结点顺序

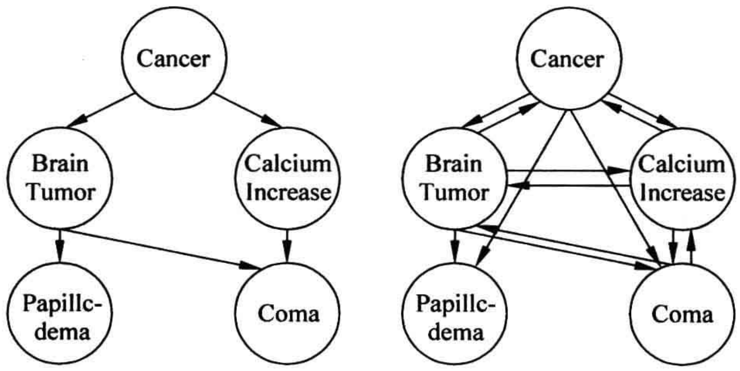
2 Compute and cache ψ functions for each X_i, X_j pair //对每一个结点对(X_i, X_j)计算 ψ 函数并缓存其值

3 For each set of candidate query variables Q //对每一组候选查询变量

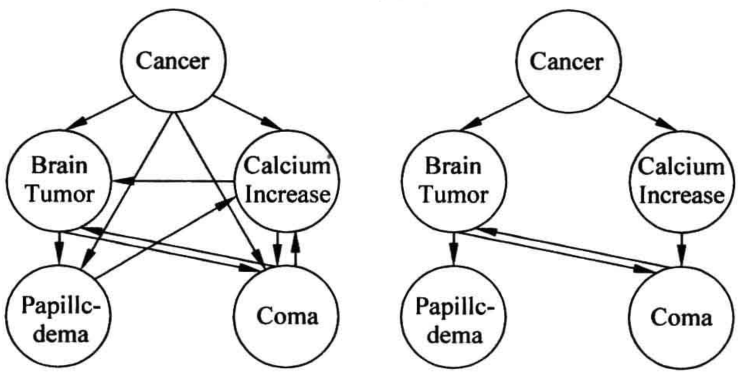
4　　For each ordering　　　　　　　　　　　//对每一个顺序
5　　　Compute the loss factor $h_{<}(Q)$ associated with the ordering with Eq.(6.37)
　　　　　　　　　　　　　　　　　　　　//使用公式(6.37)计算每组顺序的损失因子 $h_{<}(Q)$
6　End For
7　Average the loss factors $h_{<}(Q)$ obtained from each ordering to obtain the expected posterior loss factor $h(Q)$　　　　　　　//对所有顺序的损失因子求平均值,获得期望损失因子 $h(Q)$
8 End For
9 Scan expected posterior loss factor $h(Q)$ for query q with lowest value
　　　　　　　　　　　　　　　　　　　　//搜索最低损失值的查询变量组合及其取值
10 Ask query $Q{:}=q$
11 Receive complete response x
12 Update $P(G,\theta_G)$　　　　　　　　//利用获得的新数据更新贝叶斯网络参数和结构分布
13 Repeat

5. 实验结果和分析

实验对主动学习方法和随机抽样以及以均匀分布选择查询结点作了比较,每个方法对所有结点对的可能的边做评价。图 6.13(b)、(c)和(d)示出了随机抽样,均匀分布选择查询和主动查询对于 Cancer 网络[图 6.13(a)],各自估计的因果边概率。图 6.13(b)显示了随机抽样方法需要更多随机观测数据来学习一些边的方向,图 6.13(d)显示了主动学习方法比均匀分布选择查询产生了更好的结点间因果介入估计。实验反映出主动学习方法更好地揭示了边及其方向,其是唯一能在给定 50 个查询的限制下得到良好结果的方法。



(a) 真实的Cancer网络　(b) 70个观测数据以后的Cancer网络



(c) 20个观测数据和50次统一查询的Cancer网络　(d) 20个观测数据和50次主动查询的Cancer网络

图 6.13　实验结果

6.3 增量学习

6.3.1 基本原理

在某些机器学习算法里，如主动学习，最基本和最频繁的操作就是使用当前的模型在增加新的样本后对网络参数的修正。如果每次增加训练样本后都重新学习分类参数，复杂性是相当高的，然而贝叶斯学习具有增量学习的特性，使得这些计算可以增量地进行，从而大大地缩减了问题的规模，图6.14给出了贝叶斯网络增量学习的模型。约定样本空间 S 由特征空间 $\boldsymbol{X}$ 和类别空间 C 组成：$S=\{s_1,s_2,\cdots,s_n\}=<\boldsymbol{X},C>$。由于新的训练样本的加入，使得先验信息由 $p(\theta|\boldsymbol{X}_0)$ 变为 $p(\theta|S,\boldsymbol{X}_0)$。它综合了样本信息和先验信息，这正是贝叶斯增量学习模型的基础，可简单理解为

$$\text{后验知识}(\boldsymbol{X}_1)=\text{先验知识}(\boldsymbol{X}_0)+\text{样本信息}(S)$$

当新的样本到来时，上面的后验知识变成先验知识，因此它是一个利用样本知识来修正当前知识的连续的、动态的过程。

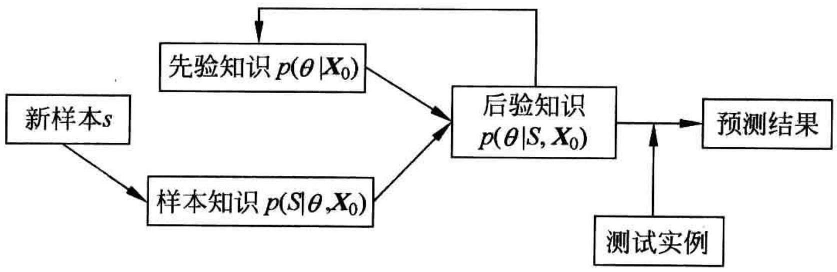


图 6.14 增量学习模型

6.3.2 贝叶斯网络参数的增量学习方法

在增量学习中，训练数据分为初始样本集 L 和新加入的样本集 L'。首先，在初始样本集 L 上进行贝叶斯网络的结构学习和参数学习，得到当前网络的结构和参数。假设模型中的结点变量为离散的，参数采用 Uniform Dirichlet 先验分布，则根据3.3.2小节中的公式(3.36)，参数估计由以下公式获得。

根结点参数估计，即

$$\theta_r=p(Y_r\mid\theta)=\frac{1+\text{count}(Y_r)}{|Y|+|L|} \tag{6.41}$$

式中，$\text{count}(Y_r)$表示初始样本集中根结点 Y 取第 r 个值的样本个数；$|Y|$表示该结点的取值个数；$|L|$为初始样本集的样本个数。

只有一个根结点为父结点的子结点参数估计，即

$$\theta_{ik|r}=p(X_{ik}\mid Y_r;\theta)=\frac{1+\text{count}(X_{ik}\wedge Y_r)}{|X_i|+\text{count}(Y_r)} \tag{6.42}$$

式中，X_{ik}为第 i 个结点的第 k 个取值。

具有两个父结点(其中一个为根结点)的子结点参数估计，即

$$\theta_{ikjt|r}=p(X_{ik}\mid Y_r,X_{jt};\theta)=\frac{1+\text{count}(X_{ik}\wedge X_{jt}\wedge Y_r)}{|X_i|+\text{count}(X_{jt}\wedge Y_r)} \tag{6.43}$$

网络中的其他参数估计可依此类推。

接下来，进入增量学习过程。增量学习过程主要是针对新增加的训练样本的处理，对于新增加的训练实例，如果当前模型(这里假设它是分类器)对其分类正确，则不更新模型结构，只进行参数增量调整。如果分类不正确，则对结构(见 6.3.3 节的“贝叶斯网络结构的增量学习方法”)和参数都要进行增量调整。

参数增量的调整公式为

$$\theta'_r = p(Y_r \mid \theta') = \frac{1 + \mathrm{count}(Y_r) + \mathrm{count}'(Y_r)}{|Y| + |L| + |L'|} \tag{6.44}$$

$$\theta'_{ik|r} = p(X_{ik} \mid Y_r;\theta') = \frac{1 + \mathrm{count}(X_{ik} \wedge Y_r) + \mathrm{count}'(X_{ik} \wedge Y_r)}{|X_i| + \mathrm{count}(Y_r) + \mathrm{count}'(Y_r)} \tag{6.45}$$

$$\theta'_{ikjt|r} = p(X_{ik} \mid Y_r, X_{jt};\theta') = \frac{1 + \mathrm{count}(X_{ik} \wedge X_{jt} \wedge Y_r) + \mathrm{count}'(X_{ik} \wedge X_{jt} \wedge Y_r)}{|X_i| + \mathrm{count}(X_{jt} \wedge Y_r) + \mathrm{count}'(X_{jt} \wedge Y_r)} \tag{6.46}$$

其中，$\mathrm{count}'(Y_r)$表示新增样本集 L' 中根结点 Y 取第 r 个值的样本个数。

6.3.3 贝叶斯网络结构的增量学习方法

贝叶斯网络结构的增量学习主要适用于贝叶斯网络的在线学习和大数据集下数据不能一次性读入内存，需要若干批次进行学习的场合。Friedman 给出了一种贝叶斯网络结构的增量学习方法，目前有多种基于此思想的学习方法的变换，在此就不一一赘述了，而只给出其基本学习策略。

最简单的贝叶斯网络结构增量学习的策略是：将数据分为若干批次来学，每一批次的学习都要找出一个后验概率最大的网络，作为下一批次数据学习的初始网络。这种算法的优点是每一次的空间开销都比较稳定合理，因为它只保存当前批次的数据，已经学过的数据完全抛弃。但它的缺点是经过若干次的迭代后，网络会锁定在某个网络模型上，丧失对新数据的适应能力。针对此问题，Friedman 提出了针对当前批次数据的学习产生一系列候选网络作为下一批数据学习的初始网络，并且引入了充分统计量 $\mathrm{Suff}(G)=\{\hat{N}_{X_i,\mathrm{Pa}(X_i)}:1\leqslant i\leqslant n\}$ ($\hat{N}_{X_i,\mathrm{Pa}(X_i)}$是结点 X 和其父结点各取值组合在原数据集中的样本数量的向量)对学过的数据集的信息进行存储保留，每一批数据学习后都为下一批次的学习做了充分的准备，保存了更多的先验知识，也使以后学出的网络能与潜在的网络模型拟合得更好一些。实验结果也表明，无论是在存储空间的压缩还是在学出的网络质量的方面，该方法都行之有效。

Algorithm6.7 给出了算法的伪代码，其中，贝叶斯网络表示为(G,θ)，F 表示由当前网络中任意两个结点之间加边、减边和逆转边等操作形成的一系列候选网络，$\mathrm{Nets}(S)$表示那些被充分统计量记录集合 S 所评价的网络结构的集合，即 $\mathrm{Nets}(S)=\{G:\mathrm{Suff}(G)\subseteq S\}$。

Algorithm6.7 The incremental learning for BN structure

```
Incremental learning(
G,                                  //初始网络结构
F,                                  //初始网络的候选网络集合
)
1  S=Suff(G)∪∪_{B'∈F}Suff(G')       //基于初始数据集，对初始网络和所有候选网络计算充分统计量
```

```
2  Forever
3  Read new data u_n                       //获得新增数据 u_n
4  Update each record in S using u_n       //使用 u_n 更新充分统计量集合 S
5   If n mod k=0 then
6          G=arg max_{G'∈Nets(S)} Score(G'|S)                 //获得评分最高的网络结构
7           Update the frontier F (using a search procedure)      //更新候选网络集合
8         Set S to Suff(G) ∪ ∪_{B'∈F} Suff(G')
9  Compute optimal parameters θ for G from S
10  Output (G,θ)
```

算法初始给定了初始网络和候选网络集 F,第 1 行计算候选网络的充分统计量;第 3 行读入一条或一批新的数据;第 4 行更新充分统计量;第 5～9 行找出候选网络中打分值最高的网络作为当前的网络结构,接着搜索找出候选网络中打分值较高的网络,作为下一批数据学习的初始网络,同时更新充分统计量;第 10 行输出最终的网络结构和相关的参数。

对于打分函数,先前的打分函数都是针对于批量数据学习的,也就是对于不同网络结构的打分都是建立在同一个数据集上,数据集是固定不变的,但是对于增量学习来说,数据集是变化的,这一条件已经不成立了。每一批数据的到来都要更新数据集,该情况也就是在不同的时刻、在不同的数据集上对不同的网络结构进行打分。为解决这一问题,引入平均打分函数,实际上就是用在批量学习上的打分函数除以记录该结点的充分统计量的结点的记录条数,定义的平均打分函数为

$$\mathrm{Score}^{*}_{\mathrm{MDL(BDE)}}(X_i,\mathrm{Pa}(X_i))=\frac{\mathrm{Score}_{\mathrm{MDL(BDE)}}(X_i,\mathrm{Pa}(X_i))}{\sum_{x_i,\mathrm{Pa}(x_i)}N(x_i,\mathrm{Pa}(x_i))} \tag{6.47}$$

式中,$\mathrm{Score}_{\mathrm{MDL(BDE)}}(X_i,\mathrm{Pa}(X_i))$为常见的批量学习算法中的 MDL 或 BDE 打分函数;$\sum_{x_i,\mathrm{Pa}(x_i)}N(x_i,\mathrm{Pa}(x_i))$为指参与对 X_i 结点打分的所有的记录条数。针对变化的数据集,该打分的含义实际上就是用平均到记录上的 $\mathrm{Score}_{\mathrm{MDL(BDE)}}(X_i,\mathrm{Pa}(X_i))$打分值来近似批量数据集下的局部分数。

第7章
概率图模型理论在计算机视觉中的应用

最近几年，概率图模型已经被广泛应用在计算机视觉领域，从低级别的图像处理任务，如特征提取、图像分割，到中级别的目标检测、跟踪和识别，乃至高级别的动作和行为建模识别，都在使用概率图模型解决这些计算机视觉问题。

本章将对近5年概率图模型在计算机视觉中的研究状况作详细的总结，重点分析了图像分割、目标检测与识别、目标跟踪、活动建模和识别中的概率图模型的应用。根据任务类别，对模型进行了分类，讨论了它们的优、缺点。

7.1 基于概率图模型的图像分割

概率图模型广泛运用于计算机视觉领域，图像分割便是其成功应用的一个典例。MRF和CRF是图像分割中常见的概率图模型，其中尤以MRF模型使用的时间最长。它们常常被用来描述图像不同区域之间的空间关系。例如，一个图像可以分解成二维的网格，每个网格点对应一个类别变量和观测变量，生成MRF模型可以用来描述所有这些变量的联合概率分布。根据贝叶斯法则，这个联合概率分布可以被分解成观测变量关于类别变量的条件概率和类别变量的先验概率的乘积。类别变量的先验概率一般是指空间相邻的图像点趋于同一类别，常常用一个马尔可夫随机场来描述。

在MRF的基础上也衍生出了很多结构更复杂的模型。例如，在文献[164,168]中，作者提出了一个多尺度分层模型。图像在不同的尺度上被分割成局部区域，结点由高层到低层对应图像中越来越精细的区域。两个相邻层中的随机变量形成一个马尔可夫链。这种多尺度分层模型可以更好地描述图像在不同尺度内的空间关系，与基本MRF模型相比，往往具有更好的效果。图7.1给出了基本MRF模型和一个多尺度的图像分割MRF模型。

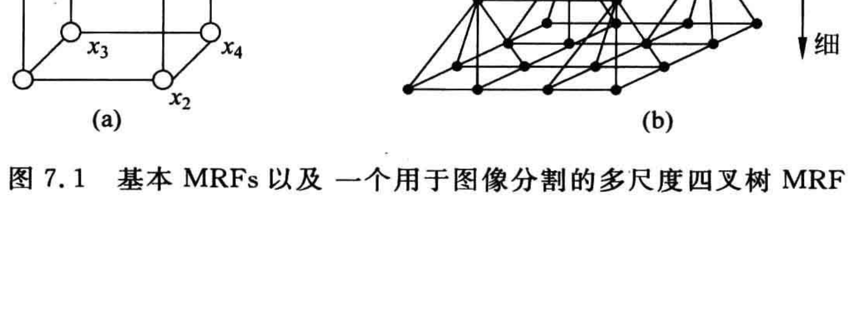


图7.1 基本MRFs以及 一个用于图像分割的多尺度四叉树MRF

多尺度随机场也被做了很多的改进。其中 Cheng 等使用类别概率树模型来表示一个更为复杂的层与层之间的条件转移概率。其中 Cheng [169]等在层间使用一个分类概率树，扩展了多尺度随机场模型，其引入了一个更为复杂的转移条件概率。这样在上层就可以使用更大的邻域(如 5×5)。由于可以使用更多的父结点，多尺度随机场能够捕捉和综合更为复杂的上下文信息。另外，Wilson 等[170]扩展了每个结点的邻域，邻域中既可以包含同层的相邻结点，也可以包含其上层的父结点。同层的链接可以描述同层结点之间的空间关系。固定结构的四叉树多尺度随机场(The fixed structure of a quadtree multiscale random field)往往会导致块状分割(Blocky Segmentation)[171,172]。为了缓解这一问题，Irving 等在文献[173]中提出了一个交叠树状模型。层中的每个结点对应的图像区域可以重叠。这样的做法使得同一层中的空间相邻的结点就有更大的可能性拥有一个共同的上层父结点，也能很好地减少不准确分割的问题。

MRF 模型基于很强的条件独立性假设，即给定一个结点的类别，这一点的观测变量与其他结点的观测变量互相独立。对于图像分割，这个假设往往无法成立。为了弱化这一假设，文献[165]中提出双重马尔可夫随机域模型(Double Markov Random Filed, DMRF)。文献[166]中提出对偶马尔可夫随机域模型(PMRF)，方法是直接假设类结点和观测结点的联合随机场满足马尔可夫特性。虽然 PMF 模型较传统的 MRF 模型具有相对比较弱的条件独立假设，但是描述类别结点和观测结点的联合势能函数非常繁琐。

CRF 模型的引入，从根本上克服了 MRF 的局限性。作为一个判别模型，CRF 允许所有的类结点和所有的观测结点之间可以有任意的连接关系。初步研究成果显示，CRF 可以成功地用于分割图像。He 等在文献[174, 175]中使用 CRF 模型来分割静态图像。通过引入额外的隐结点，他们还能够综合地利用上下文信息(如图像的背景环境信息)。他们的实验结果显示，较之 MRF 模型，CRF 能够更好地进行图像分割。在文献[176]中，Ren 等使用 CRF 模型来进行图像标注。通过定义额外的能量函数，他们在 CRF 模型中整合了多种信息[如相似性、连续性和亲密性(Familiarity)等]。但这个模型的缺点在于，它用一个层来连接所有的隐结点，而实际上隐结点可能表示多层结构化的信息。

CRF 模型也能够处理更加复杂的图像结构。文献[177,178]中提出一个多层树状 CRF 模型，对不同的尺度层中的类别结点加上一致性的约束。在文献[179]中，Toyoda 等提出了能够综合利用图像局部及全局空间关系约束的 CRF 模型。其中全局关系通过一个全局互势函数(Pairwise Potentials)来实现，这些全局是函数能充分描述全局的场景信息。这种自上而下的全景上下文信息，可以有效地校正基于局部图像的标注，显著地提升标注效果。图 7.2 显示了一个用于目标检测和分割的 CRF 模型。在传统的 CRF 模型基础上，它引入了一个附加的隐结点层，隐结点层$\{h\}$用来表示目标物的不同的组成部分的类别以及这些组成部分之间的相互关系，而这种相互关系是有助于图像标注的另一非常重要的信息。

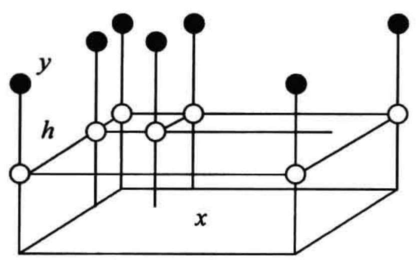


图 7.2 一个用于图像分割的隐随机场(HRF)模型

对于图像分割，贝叶斯网络模型能被用来表示关于图像分割中区域、边缘和观测值等不同元素之间的各种因果关系的先验知识。在文献[181]中，Mortensen 等基于一个两层的贝叶斯网络提出了一个半自动的图像分割技术。通

过贝叶斯网络来构建边缘和角之间的关系，对图像分割加上了一些简单的连接性的约束。给定一个用户输入的种子路径，他们使用最小路径生成树（Minimum-path spanning tree Graph）搜索发现最可能的目标边界。在文献[182]中，Alvarado 等使用贝叶斯网络模型来捕捉一个场景中人手持物体的图像分割的全部可获得的信息。这个模型可以将图像中的高层信息，如可能的人手位置，和底层的图像观测信息结合起来，推测属于人手区域的概率。Zhang 和 Ji [298]等引入了一个多层贝叶斯网络模型用来描述图像的区域，边界和角度之间的结构化的因果关系。这个模型既能利用物体形状的局部约束（如平滑型和关联性），又能利用全局约束，同时也能有效利用用户输入信息。Feng 等[184]把贝叶斯网络和神经网络结合起来用于场景分割。在这个应用中，神经网络用于产生图像类别的尺度概率，而一个固定结构的四叉树状贝叶斯网络模型（BN-quadtree Model）则用于表示类结点的先验分布。图 7.3 示出了这个多尺度贝叶斯网络模型的结构。

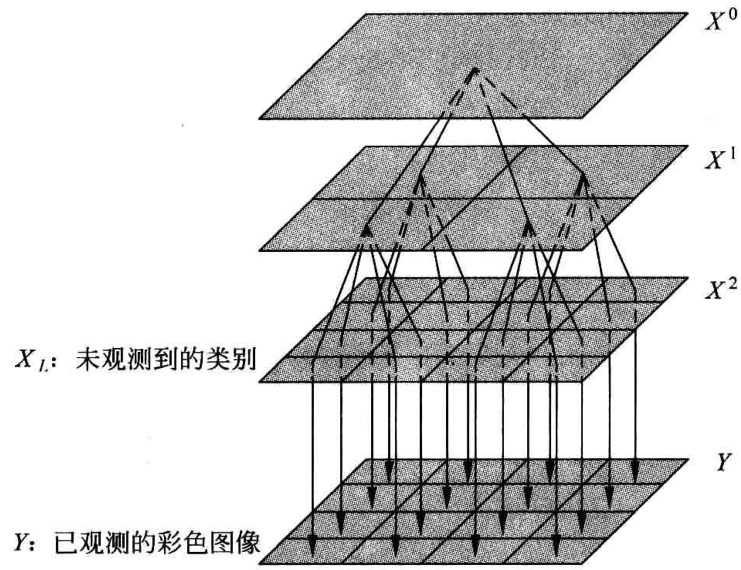


图 7.3 一个用于图像分割的多尺度四叉树 BN 模型

为克服四叉树贝叶斯网络的固定结构带来的缺陷，Todorovic 等[172,185]，提出了一个动态的多尺度树状贝叶斯网络模型。他们的方法能够同时获得最优化的结构（即在同层相邻结点间的连接）及每个结点的最优值。虽然实验结果显示了该算法的可行性，但是这个模型过于复杂而且需要做大量的随机变量推算。

所有的这些相关工作，充分显示了贝叶斯网络模型具有很强的能力来整合多源信息，并用于矫正和提高图像分割或标注的效果。

7.2 基于概率图模型的检测和识别

与低级别的图像分割任务相似，概率图模型已经成功应用到目标检测和识别中。图像目标检测是发现特定的目标（如人脸）在给定的图像中是否存在，而图像目标识别是识别对象的特定实例（如乔治・布什）。对于目标检测，AdaBoost 算法[186]是一个被广泛使用的方法，它选择少量的重要的视觉特征但却输出多个高效的分类器。最近词袋模型（Bag of Words）被成功应用在目标识别问题上，其目标模型的学习是通过建立一个局部属性的码书

(Codebook),该方法忽视了目标的几何信息。虽然这些方法能够很好地进行目标识别,然而,它们却要求大量的训练数据,并且在学习过程中不容易融合先验或相关信息。由于训练过程中没有使用目标的几何(Geometric)信息,一些与目标本身无关的信息可能会被加入,导致模型的过拟合(Over Fitting)。然而概率图模型在目标检测和识别中的应用可以克服这些问题。

使用MRF模型,目标检测和识别可看作是一类图像标注问题[126]。DasGupta等[187]提出了基于MRF模型从模糊不清的图像中同时复原和识别目标物体的方法。他们引入了部分消息传递(Partial Message Propagation)的概念,任何给定的图像分块(Patch)只部分受近邻影响,并且依赖于两个近邻间的空间交互关系。识别过程发生在不断循环的复原阶段,复原阶段帮助其减少搜索空间。基于MRF的识别技术在给定手工初始化时,取得了相当好的结果。然而,由于缺乏清楚的形状模型,这些方法在自动目标识别和图像分割应用中往往会失败。Kumar等在他们的ObjCut[188]工作中,提出了一个用于目标检测的贝叶斯方法,这个方法使用图示结构表达(Pictorial Structures (PS) Formulation)和MRFs模型一起提供物体形状的先验知识。形状模型(Shape Model)提供了一套潜在的变量,这些变量将倾向于某种物体形状的分割。对于表情识别和动作识别,Zhang[189]和Tong等[190]提出了使用DBN模型捕捉在人脸表情(动作)与图像观测值之间的(Spatio-temporal)空间-时间关系。虽然形状先验知识能够帮助减少分割错误,但是当物体偏离标准形状模型时,形状模型也引入了分割误差。

为了识别可变形的目标物体,生成式星座图模型(Generative Constellation Models)被引入到形状的建模过程中。这个模型首先由Burl等[191]提出,在这个模型内,目标物体由在可变形的结构(Deformable Configuration)中的特征部分所表示。相似的,Weber等[192]提出了将一个目标类(Object Class)表示为一类由刚性部分群组成的可变星座图模型(a Flexible Constellation of Rigid Parts),其中目标类被定义为具有相同特征属性部分群(Parts)的集合或相似且在相似空间结构存在的部分群集合。星座图(Constellation)模型有几个优点:模型清楚地表示了形状的变化和由于Occlusion和检测器误差而引起的属性存在与否的随机性。这个模型被Fergus等扩展到表示外观的变化[193],他们扩展了[191,192],利用该模型从有大量形状变化的目标物体训练数据集中进行有效学习。Li[194]提出了一个增量学习贝叶斯网络的方法,其从101个变形目标物体类别中学习可视目标,每一类使用了少量的图像训练数据。Loeff[195]等对于目标检测和定位提出了一个生成式概率模型,其可以用最小监督方法被建立起来。他们的临时星座图模型(Improvised Constellation Model)能够处理Occlusion和在属性特征抽取中的误差问题,并且使得模型以图像中的部分群数量和属性数量级进行线性推理。

除了生成式模型,判别PGM模型也已经被广泛应用在目标检测和识别中。对于在图像中检测人造结构的问题,判别随机场(Discriminative Random Fields,DRFs)被用来建立类别变量之间高层次上下文关系。关联$A(X_i,Y)$和交互势能(I_{ij})被引入进来,表示近邻区域的观测和交互。在DRFs框架中,$A(X_i,Y)$用Logistic Regression判别模型建立,而I_{ij}作为代价函数用来惩罚每一对相异的类别。这种交互形式有益于对离散的观测数据进行类别平滑分段。然而,由于DRFs中交互势函数的额外计算量,使得其比MRF计算更为复杂。Quattoni等[196]对于基于CRFs的目标识别提出了一个判别模型,其条件分布$p(y|x)$是直

接建模得到，推理和参数估计使用标准的可信度传播算法（Belief Propagation）得到。在文献[180]中，作者提出了一个复杂的 CRF 模型用于目标检测。他们引入一个额外的隐含层来表示被检测目标各部分的位置。目标各部分与目标中心位置有约束关系。基于相似的模型，Winn 等进一步在各部分中引入了布局（Layout）一致关系，用以在二维和三维图像中[198,199]识别和分割部分 Occluded 的目标。Kumar 等在文献[199]中提出了一个两层的 CRF 模型，这个模型对于目标类别分类既能够捕捉像素对的上下文信息，又能够捕捉长范围上下文信息。

7.3 基于概率图模型的目标跟踪

目标跟踪任务是对视频每一帧中的一个对象或多个对象（多目标跟踪）进行定位。不同于目标检测，目标跟踪充分利用了连续视频帧中物体平滑运动的特点。基于这个假设，目标跟踪算法减少了发现非目标的概率，并且限制了在未来视频帧中的检索区域。尽管在过去数十年间，目标跟踪在计算机视觉领域得到了深入的研究，但它仍然是一个具有挑战性的课题[222,223]。

从概率角度来看，跟踪是一个（在线）实时滤波问题，对 $t=1,2,\cdots$，估计 $P(x_t \mid y_{0\cdots t}, x_0)$，这里 x_t 是对象在时刻 t 的状态（假设初始状态 x_0 给定），包括位置、姿势、尺寸、旋转等，y_t 是 t 时刻的图像帧。在视觉跟踪中，基于 PGM 的方法往往使用动态模型（如一阶状态—空间模型），描述（$\{x_t\}$，$\{y_t\}$）的联合分布，如图 7.4 所示，从而把跟踪问题简化为状态—空间模型的推理[224]。

1. 关节式链物体跟踪

在关节式物体跟踪（Articulated Object Tracking）领域，概率图模型受到了极大的关注。最常见的一个例子是人体跟踪。人体跟踪就是基于视频中的图像特征，估计出二维或三维的人体姿势（身体各部分关节的角度和位置）。

在文献[224]中，切换动态线性系统（SLDS）被用来刻画诸如人体形态运动这样复杂多样的动态行为。为了处理动态变化，在状态变量之上引入了离散切换变量 $S=s_1,\cdots,s_T$，如图 7.5 所示。在实际应用上，SLDS 模型的推理[即 $P(S,X|Y)$]只是近似的。文献[225]介绍了 3 种不同的近似方法。维特比（Viterbi）近似利用了 $P(S,X|Y)\approx P(X|S^*,Y)$，在维特比算法中，$S^*=\arg\max P(S|Y)$。一种变化推理算法将 SLDS 模型分解为上层的隐马尔可夫模型（HMM）和下层的线性动态系统。最后一种归纳伪贝叶斯算法试图将模型拆解成较少（固定数目）部分的混合物。SLDS 模型表现出优于简单动态模型的跟踪稳定性，在处理包含多种类型的混合运动时也很有效。

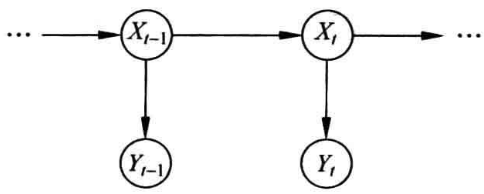


图 7.4 用于目标跟踪的一阶状态—空间模型

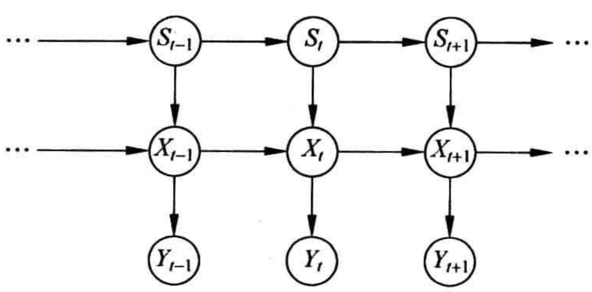


图 7.5 切换线形动态系统

人体是一个典型的非刚性关节式物体(Articulated Object)的例子。关节式物体由连接物和关节构成的,因此使用图模型可以自然地表达人体和姿态检测时的关联制约关系。自从早期的工作[227,228]开始,各种生成式概率图模型就已经受到研究者们的欢迎,并且复杂程度逐渐增加[229,230,231],包括处理姿势和视角变化的能力[232]。在文献[226]中,人体被表示成一个松散连接的肢体集合。通过双马尔可夫模型(Pairwise MRF)(图 7.6)建立肢体三维姿态(如身体各部分的位置和方向)的条件独立关系,其中边缘势函数(the Edge Potential Function)(空间或时间上)被定义为一个高斯混合模型,它的参数通过 EM 算法从基准运动的捕捉数据中获得。模型通过非参数信息传播(即通过采样粒子数据近似消息)进行推理[233]。这种方法的主要优点是利用从各身体部位检测器获得的自下而上的信息,将搜索空间的复杂性减少到肢体数目的线性空间。对于上半身跟踪问题,Zhang 等[234]提出了用动态贝叶斯网络建立所有可能的身体动作。他们在模型中引入了各种上肢动作的通用物理知识和解剖学的关联和限制关系,其不同于大多数现有的上肢运动模型,该模型可以跟踪身体任何一种动作,而不只是一些典型的动作。

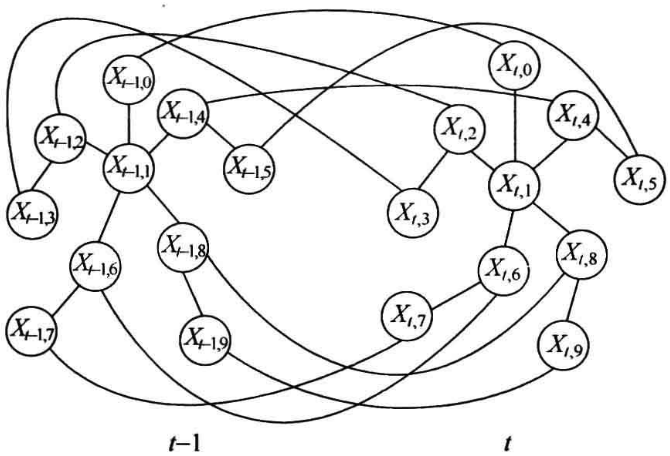


图 7.6 用于肢体建模的图模型

2. 判别式的目标跟踪方法

不同于生成式的目标跟踪方法,判别式的跟踪方法直接优化跟踪器的预测精度,即 $P(X|Y)$,这是最可取的方法。Ross 等[237]提出了一个具有隐变量的条件模型,这些隐变量可以进行灵活地切换某个特征属性。该判别模型允许某些可以提高准确性的相关属性特征包含到模型训练中来。这种方法已经成功应用于复杂轨迹的目标跟踪。Taycher 等[238]为了应用一个类条件随机场 (CRF-like)判别模型进行目标跟踪,把连续状态空间离散成网格。提出的网格滤波算法几乎可以实时进行准确的目标跟踪。然而,这种基于网格的方法通常需要大量的姿态数据作为先验知识以实现良好的近似。

文献[239]提出的有向条件概率模型已经被扩展到连续空间的三维姿态估计问题[235],基于图 7.7(a)所示的图形结构,其条件分布被分解成局部预测器 $P(x_t|x_{t-1},y_t)$。他们通过先验混合贝叶斯网络建立局部预测器,因此命名为混合贝叶斯探测器马尔可夫模型(Bayesian mixture of experts Markov model,BM³E)。他们对模型的时序推理提出了一个混合密度传播框架(Mixture Density Propagation Framework)。

除了判别模型,另一种提高跟踪准确性的方法是判别地学习生成式模型。Kim 和

Pavlovic[241]已经证明,即使像简单的线性模型 SSM,通过复杂的判别式学习也能大大提高跟踪精度。这里有两种学习算法被提出来进行判别地学习生成式模型:条件最大似然(CML)和 Slice-wise 条件最大似然(SCML)。不幸的是,两者的学习目标函数都不是凸函数 Convex,导致模型对初始参数和优化过程极其敏感。为了解决上述问题,一种新的无向判别图模型——条件状态空间模型(CSSM)[图 7.7(b)]被提出来[236]。CSSM 在性质上类似 CRF,在整个测量序列上以状态为条件,同时利用问题的序列结构。CSSM 的参数学习是一个带有凸可行约束的凸函数优化的例子,以增强密度可积性。他们提出的推理算法远远快于卡尔曼滤波,同时允许模型有大量的测量属性。

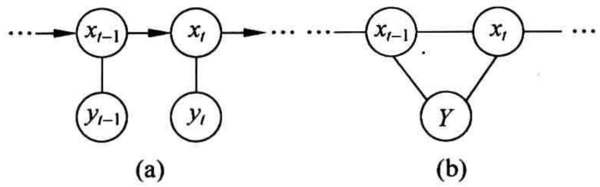


图 7.7 BM³E 和 CSSM 模型

3. 多目标跟踪

多目标跟踪的目的是定位未知数量的(可能是随时间变化)多个对象。多目标跟踪通常比单目标跟踪更具挑战性,因为需要同时估计跟踪器的状态和跟踪器测量值的关联(Measurement Association)。在基于概率图模型的多目标跟踪方法中,除了状态变量和观测变量外,还存在跟踪器的关联变量,其辅助跟踪器进行特别的观测。例如,Khan 等在文献[242]使用了文献[243]中的联合粒子过滤器(Joint Particle Filters),其间目标物体的交互是通过 MRF 模型建模,该 MRF 模型的势能函数计算两个跟踪器之间重叠像素的数目,以避免物体平压变形,这是从如昆虫这样的群居生物中观察到的性质。最近,Yu 等[244]提出了一种联合表示跟踪状态变量和关联变量的概率图模型。该模型可以分解成关联先验模型和似然模型。前者是由具有异或指标器(the Exclusive-OR Indicator)势能的双马尔可夫模型(Pairwise MRF)建模。这个框架具有很强的计算能力,因为它允许学习过程以分布式的方式完成。多人脸特征跟踪是另一个有趣的应用。经过个体特征跟踪以后,Su 和 Huang[245]使用了信息传播来推断空间域中的每个人脸特征轮廓,每个轮廓中的人脸特征关系是从拥有大量人脸表情数据的数据库中学习得到的。

7.4 基于概率图模型的行为建模和识别

从静态的概率图模型到不同类型的动态概率图模型都被广泛用到人类行为建模问题。在本节中,回顾不同概率图模型的行为建模技术,并比较行为识别中的静态和动态模型。

1. 静态模型

贝叶斯网络及其相关变型常常被用来描述某个行为若干基本元素之间的层次关系。典型地,一个 3 层 BN 被用来描述图像特征、基元行为以及复杂事件之间的关系。对于场景识别,Hongeng 等[246]提出了一种分层 BN(Hierarchical BN)来描述图像特征、子场景和主场

景之间的关系。对于动作识别，Filipovych 等[247]使用一个 BN 模型来描述表演者和物体之间的关系，如图 7.8 所示。该模型包括 4 种类型的结点，即时序状态结点(M)、静态状态结点(S)及其相应的观测值结点(O_M 和 O_S)。

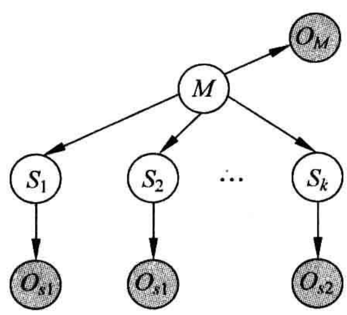


图 7.8 表演者-物体贝叶斯网络模型结构

各种贝叶斯网络诸如 Laten Dirichlete Allocation (LDA)、Hierarchical Dirichlet Process (HDP) 和 Probabilitic Laten Semantic Analysis(PLSA)都已被广泛地应用到自然语言处理中。由于人的行为结构可以很容易地用语言处理中的概念进行表示，如基元动作类似于视频单词(Word)、动作类似于子主题(Sub-topic)、行为类似于主题(Topic)，因此，这些模型能够自然而然地被扩展到行为建模和识别中。Wang 等[248]利用 LDA 和 HDP 建立模型来描述视觉特征、基元行为与行为这些不同层级的行为之间的关系(利用基元行为之间的交互)。不同层次之间(如特征与动作之间、动作与行为之间)的定量关系采用多项分布进行描述。Wang 等在文献[249]中也提出了类似的模型用于识别未校准多摄像头视频中的行为。PLSA 已被用来分析一个动作中同现的基本动作，一项行为中的同现动作。Niebles 等在文献[250]中提出了一个用于行为识别的 PLSA 模型，其层次结构分为 3 层：特征属性(单词)、动作(潜在主题 Latent Topic)和行为(主题)。而文献[251]中提出了一个稍微不同的方法，其中间层描述人体的不同部位(而不是基本的动作)，而比它低一层的是身体部位的量化特征。

2. 动态模型

(1) HMM 及其变型

继成功地被应用到语音识别上以后，HMM 又被广泛用于视频行为建模。HMM 中的隐结点状态变量代表一个动作分解而成的基元动作，而隐结点的转移概率代表了各个动作的动态变化概率。观测结点提供了基元动作图像属性的测量值。Zhang 等[252]提出了一种半监督 HMMs 用于非寻常事件的检测。这个模型分为两个层次，其中基元事件(动作)由若干具有最小持续时间约束的子隐马尔可夫模型建模；上层活动被建模为一个遍历的 K 类 HMM 模型，其中隐状态是基元事件(动作)。在 Ivanov 的工作中[253]，低层次的基元行为(动作)的检测也使用了一个标准的 HMM 模型。

为了克服基本隐马尔可夫模型的局限性，研究者提出了一些新的 HMMs 模型及其变型。为了解决图像特征的高维度问题，多观测变量的隐马尔可夫模型(MOHMM)[254][如图 7.9(b)所示]被提出来以分解观测空间，该模型假设在给定隐状态变量的条件下，不同的观测结点变量是相互条件独立的。通常这需要对图像特征进行仔细的检查，以确保该项假设对于行为识别是有效的。为了表示多个实体之间的交互关系，Oliver 等在文献[255]中使用了耦合的 HMM(CHMM)建立实体之间的交互关系。正如图 7.9(c)所示那样，CHMM 包括了几个 HMMs，耦合的隐状态结点代表了对象之间的相互交互。不同于 CHMM 模型，Xiang 等[256]提出了动态多链接隐马尔可夫模型(Dynamic Multi-Linked-Hidden Markov Model, DML-HMM，如图 7.9(d)所示)，该模型对一个动作中的多目标交互关系进行建模。DML-HMM 和 CHMM 都是通过各组成 HMM 之间结点间时序的链接来耦合各组成

HMM。DML-HMM 和 CHMM 的主要区别是，DML-HMM 可以自动发现必要的各组成 HMM 之间的交互链接，而 CHMM 则进行完全的耦合链接，所以，DML-HMM 减少了冗余链接，提供更优化的状态空间分解。这使得 DML-HMM 对群体动作和行为建模时具有更可控的计算量。此外，由于行为建模的观测数据往往带有噪声，一个具有较少参数的状态空间分解能在训练阶段更好地对统计分布进行估计。

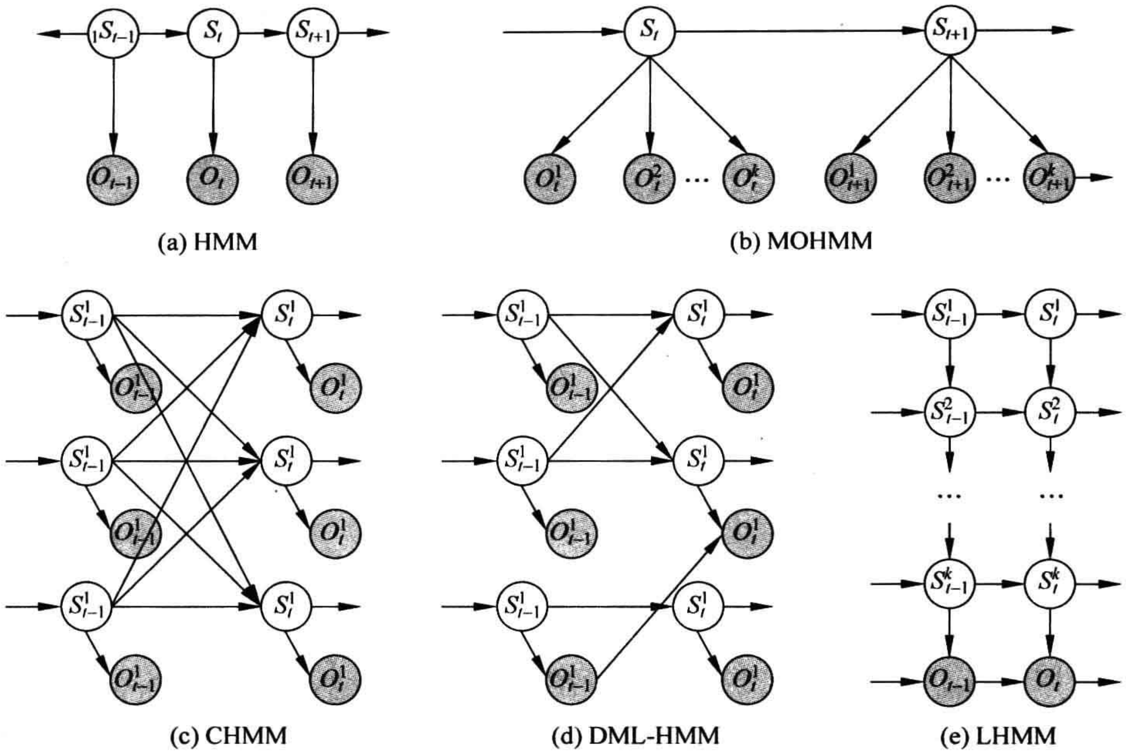


图 7.9 基本 HMM 及其变型

为了在一个行为模型中表示多种行为，HMM 的不同扩展模型被提出来。层次化的 HMM 模型(LHMM)，如图 7.9(e)所示，包含多层并行运行的 HMM 模型，低层产生高层的观测值，底层是图像的观测值。Oliver 等在文献[257]中第一次运用了两层的 LHMM 识别多级行为。该模型上层构建行为中构成动作的转移关系，底层构建基元动作之间的转移关系。开关隐半马尔可夫模型(S-HSMM)如图 7.10 所示，是由 Duong 等提出来[258]进行多层行为识别的，可以被视作 LHMM 的扩展。其底层使用隐半马尔可夫模型来表示基元行为及其持续时间；顶层表示一个高层次的行为序列，这些高层次的行为是由基元行为序列构成。因此，可以把顶层隐状态看作是底层模型的开关因子。层次化(Hierarchical) HMM (HHMM)在多级行为识别中也是一种流行的整合动作模型和行为模型的隐马尔可夫模型。在 HHMM 中，每个状态变量被认为是自包含(Self Contained)概率模型。更精确地说，每个状态自身就是一个 HHMM(或 HMM)。这意味着，HHMM 的状态包含序列的观测变量，而不是如标准的 HMM 那样包含一个观测变量。Lv 等[259]使用动作网络(Action Net)来构建多层的行为模型，这种建模方式显示出了动作网络的状态转移图，因此该模型本质上仍是一个层次化 HMM。在动作网络中，Viterbi 路径搜索方法被采用来推测视频序列中的关键姿态、动作和行为。

最后，Hongeng 等[246]用概率有限状态机(Probabilistic Finite-State Machine，PFSM)建立子场景(动作)和场景(行为)之间的关系，该模型可以处理交叉动作的识别。改进的

Viterbi 方法用来推理场景。为了处理同样的问题，Shi 等[260]提出了一个传播网络(Propagation Networks)来表示人类行为，该模型实际上是一个 FSM 的多层次化扩展。

(2) 广义动态贝叶斯网络

所有的 HMM 模型都可以被视为特殊的动态贝叶斯网络(DBN)。虽然 HMM 的某些变型模型可以表示行为识别中不同对象之间的某类交互关系，但是这些交互关系是被这些 HMM 变型模型的特定结构所限制的。相比之下，动态贝叶斯网络能够表示不同行为实体之间更广义的因果关系，这些关系既可以是在同一层次之内的，也可以是在不同层次之间的。Wu 等在文献[261]提出了一个 DBN 模型，其结合 RFID 和视频数据联合推断最可能的行为和目标对象类别。如图 7.11 所示，该模型包括 3 个层次：行为级、目标对象级和观测级(RFID 和视频观测值)。然而，该模型没有考虑同一级不同实体之间的关系。

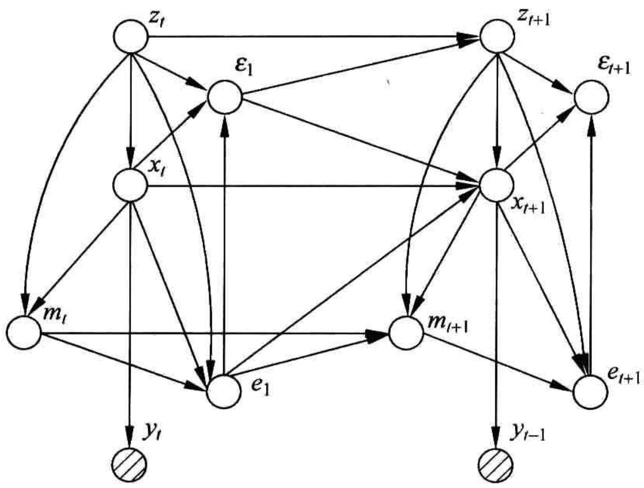


图 7.10 开关隐半马尔可夫模型

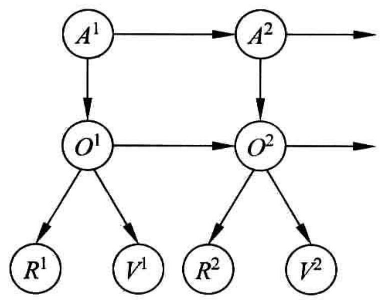


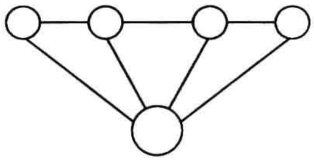
图 7.11 用于行为识别的 DBN 模型

Laxton 等在文献[262]中的 DBN 模型使用层级结构，同时表示不同层级之间和同一层级内的关系。每一项行为是由一个基元行为序列构成，而其中一些基元行为可以分解成一个基本体(Primitives)集。一种基于目标物体的层次结构被用来观测基本体及其他不可分解的基元行为的各种状态。该模型没有采用隐状态来代表每一个动作，而采用一个隐结点来表示一个基本的动作。通过这种方式，他们可以更加清楚地建立不同动作之间的交互关系。

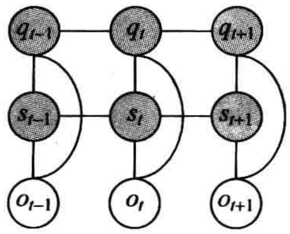
(3) 判别式模型

除了生成式(Generative)的概率图模型如 HMM 和 DBNs，判别式(Discriminative)的概率图模型也被应用到行为建模中。Vail 等[263]采用了线性链 CRF[图 7.12(a)]实现动作识别。作为与标准 HMM 模型相对应的判别式模型，线性链 CRF 在特征属性不满足相互独立性假设时显示了其优势。为了在模型中表示多目标物体间的交互关系，Wang 等[264]提出了因子 CRF(FCRF)模型。如图 7.12(b)所示，FCRF 有多条由隐结点构成的线性链，不同链上同一时刻的结点之间相互连接(FCRF has multiple linear chains of hidden nodes with connections between cotemporal nodes)，这允许 FCRF 表示分布式的隐结点及其它们之间的相互作用关系。Shi 等[265]提出了一个半马尔可夫模型(Semi-Markov Model)来建模人的动作。半马尔可夫模型本质上是条件随机场(CRF)的扩展。上述 CRF 模型及其扩展都只是限于单级行为(动作)的识别，Truyen[266]等建出了增强的马尔可夫随机森林

(Markov Random Forests,MRF)并将其应用于多层次行为识别。他们使用的模型实际上是一个隐动态条件随机场(Hidden Dynamic Conditional Random Field ,DCRF)(图 7.13)。该模型底层建立基本动作模型,上层描述高级别的行为。Boosting 算法被采用来学习势能函数中的参数。



(a)



(b)

图 7.12 线性链 CRF 和因子条件随机场

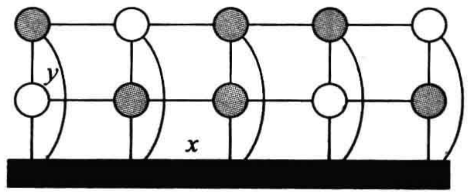


图 7.13 隐动态条件随机场

第8章 贝叶斯网络在电信个人信用风险分析中的应用

信用风险评估是经济领域中的一种分类决策问题。随着我国社会经济的发展，信用危机在各行各业日益凸现。尤其近年来，在电信领域，每年电信用户欠费的增长率高达20%，远高于收入的增长率，用户欠费造成的通信部门损失超过200亿元人民币，严重影响了电信企业正常经营和发展。虽然全国各级电信公司先后建立了用户话费和信用额度控制与管理的相应机构，但是，真正对客户信用风险进行深入地分析并提供科学、准确的评估模型却很少，某些地方还出现了一刀切和名不副实的现象，给信用良好的用户带来一些不便，致使客户满意度下降和用户的流失。因此，我国的电信运营业迫切需要进行用户信用评估研究，建立科学合理的评估模型。

但是，由于目前我国的电信客户信用评估工作处于起步阶段，评估体系和技术方法远未成熟，大量带有信用风险类标签的历史客户数据并不存在，因此通过传统的分类算法建立信用风险评估模型来预测新客户的信用风险并不可行。目前，为了评价客户的信用风险，只能由领域专家手工评价、标注客户数据，这是件极其耗时和代价昂贵的事情。

基于以上背景，本章介绍如何应用主动学习的贝叶斯网络建立客户信用度评分模型、客户信用等级评估模型，以及应用基于半监督主动学习的动态贝叶斯网络来建立客户信用风险行为演化模型。实验结果证明，基于主动贝叶斯网络的信用风险评估模型准确率高，使用的带类标签训练数据少，这有效解决了人工评估和标注客户数据的工作量，并为客户的信用评估机制和市场营销决策提供了科学的依据。

8.1 电信客户信用风险评估概述

在电信企业的决策支持系统中，以中国移动为例，用于评价客户欠费风险的指标称为信用度，信用度通常是0～100之间的连续值，以便对客户信用状况进行量化。信用度已成为进行客户消费情况分析、客户定位分析、客户风险系数分布分析、经营风险分析及客户欺诈行为分析的一个重要指标。但由于目前缺乏合适的信用度计算算法，客户的信用度主要还是凭借经验人为分配。

北京邮电大学计算机科学与技术学院的张玉洁教师针对客户申请时的初始信用度计算作了研究。武汉大学电子信息学院的徐继生协同广东移动珠海分公司的王道恒等对电信客户的信用度分配作了研究，他们都采用了遗传算法和多元线性判别，只是各自选择的评估指标有所区别。

在这些研究工作中，有采用客户的基本信息如性别、年龄、职业、学历、收入、婚姻等静态

属性对应某个规则进行分配取值，有的除了客户静态属性外还使用了一些用户的行为属性，某个客户的这些属性值之和即是其信用度，即

$$C = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \alpha_{ij} \delta_{ij}(x, k_i) \tag{8.1}$$

式中，C 为用户信用度评分；α_{ij} 为第 i 种属性的第 j 种值的权值；N 为共选有 N 种属性；M_i 为第 i 种属性共有 M_i 种可能的量化取值；k_i 为决定某个用户第 i 种属性的取值；δ_{ij} 为一个阶跃函数，其取值为

$$\delta_{ij}(x, k_i) = \delta_{ij}(x - k_i) \begin{cases} 1, x = k_i \\ 0, x = 0 \end{cases} \cdot \tag{8.2}$$

通过遗传算法的计算，能够得到多组解（图 8.1），供领域专家参考和进行选择，以解 2175431 为例，解释一下其信用度分配值的含义。

1276543，1376542，1376543，2365431，1276542，2176541，237542，2376541，2175431，
3174321，3176541，3276431，3175421，3165431，3175431.

图 8.1 遗传算法计算得到的各属性取值的多组解

若属性 1 是性别，属性取值分别是男、女，则对应的信用度分配值 Value[1,1]＝2，Value[1,2]＝1. 属性 2 是学历，属性取值分别是研究生、大学本科、大专、中专或高中、其他，对应的信用度分配值是 Value[2,1]＝7、Value[2,2]＝5、Value[2,3]＝4、Value[2,4]＝3、Value[2,5]＝1。根据这个解，利用初始信用度计算公式得出任一位客户的初始信用度。若一客户是男性，研究生，则该客户的初始信用度为 9。

那么以上的研究在一定程度上克服了信用评分过程中属性权值分配的主观性。但存在以下几点问题：

1）信用度计算公式仍然有很强的主观性

从他们的研究结果来看，信用度的计算属于演绎信用评分，仍然是半客观的，获得权值是客观的，但是属性值的权值分配仍然是主观决定的，即哪些解对应哪些值。

2）评分指标集很不完善

对客户静态信息与客户信用风险的相关性没有进行具体的统计分析，使得评估指标集存在很多无关和多余的指标参与计算，造成所建模型使用的数据量大、训练时间长、模型复杂。另外，在进行信用度计算时虽然从 100 多种用户行为属性中遴选出 10 余种重要属性纳入预测和评估，但经遴选发现其行为属性还很不充分，补充某些属性或属性组合是必要的。因此，需要对信用评估指标体系进行相关性分析，以便找到与信用风险相关性密切的属性参与建模运算。

3）信用评分模型准确率不高

由于线性模型的局限性，不能很好地逼近实际情况，有必要采用更精确的模型。

4）对于客户不同发展时期的信用评分模型没有区别对待

客户不同发展时期的信用评分的指标与计算模型是不同的。例如，新客户的信用评估指标与老客户的信用评分指标就有所区别，其指标之间的关系也有可能不同，因此，两个模型应分别建立。

5）缺乏对客户信用风险的定性分析

目前的信用度评估工作集中于对信用度的定量分析，但是通过对金融领域客户信用评

估的认识认为,为了更好地进行信用决策和风险管理,需要对客户信用风险状态进行评级区分。通过信用评级,有助于对客户群行为特征的区分,更好地预测客户未来的信用风险及企业的损益程度,从而建立预警机制。再完美的信用风险模型也仅仅是进行信用风险度量和管理的工具,任何复杂的数量分析都不能代替风险管理中的经验判断,因此有必要对客户信用进行定量与定性相结合的研究。

6) 应采用基于历史数据的实验信用评分方法进行信用度评估建模

从单个用户的角度来看,各种用户行为属性似乎是千差万别、彼此没有联系的。但从统计学的角度来看,不同类型用户行为属性的特征是有关联的,存在统计意义上的规律性。为了能够区分用户的种类,就需要找到一些用户行为属性及其组合,达到一定标准的用户是高信用度(非欠费)用户,达不到标准的则是低信用度(欠费或可能欠费)用户。用户的信用度受多种因素的影响,目前尚不能对这一问题在理论上做出圆满的解释。处理这一问题,通常采用数理统计学的方法,即把信用分析看成是数据挖掘中的一种分类问题。在目前的理论状况下,它可能是最有效的方法,也是国际学术界视为主流的方法。

考虑到信用评估问题实际上是数据挖掘中的分类问题,那么鉴于以下几点情况,贝叶斯网络理论作为信用评估建模的分类算法是非常合适的。

- 电信客户信用风险状态具有随机性和不确定性。电信客户的行为状态存在着很大的不确定性,其出现和变化不能通过简单的物理规律加以确定。例如,给定两条客户数据,它们看起来极为相似,但事实证明,一个客户会发生欠费行为,而另一个则不会,客户欠费与否具有不能直接测量的随机性。客户的拨打行为、缴费情况、业务使用情况也同样存在着不确定性。而贝叶斯网络理论在处理不确定性方面有其独特的优势。
- 信用风险评估模型的可解释性对于信用决策来说很重要。贝叶斯网络较其他算法具有简洁、准确和易理解的特点,并且其结点间的因果关系有利于信用决策分析。
- 客户不同的评分模型具有不同的动、静态特点。例如,客户申请模型属于静态分类问题,而客户行为评分模型则属于与时间相关的动态分类问题。贝叶斯网络及其扩展——动态贝叶斯网络,能够建立客观世界的静态模型和动态模型。
- 信用评估应采用定量与定性分析相结合的方法,贝叶斯网络理论是描述变量之间定性与定量依赖关系的图形模式。贝叶斯分类并不把一个对象绝对地指派给某一类,而是通过计算得出属于某一类的概率,具有最大概率的类便是该对象所属的类。在应用中可以取客户属于高信用度类别的概率作为信用分数,概率是 0～1 之间的连续数值,而电信客户信用评分的分数是 0～100 之间的连续值,为方便表示,将概率扩大 100 倍,作为最后的信用度,表示客户信用风险的大小,分数越高说明欠费风险越小,信用越好。而客户具体属于哪一类是通过其属于不同类的概率的大小比较而确定的。因此,通过贝叶斯网络建立的信用评分模型有两个输出结果来表示客户信用风险,分数是客户信用风险的定量度量,而类别是定性尺度。这种方法克服了特征属性权值分配和划分类别的临界值确定的主观性。
- 鉴于我国电信行业开展信用评估工作刚刚开始,对客户信用状态的评价只能由专家评估,分类成不同的风险水平,并由手工标定完成,其工作耗时而且代价昂贵,因而要想获得大数据量带有信用类别标签的数据是一件很困难的事情。这就要求分类算法能够利用专家知识并结合数据进行建模,同时也要求分类算法在训练样本数据少的情况下能够准确建模。而贝叶斯网络能够使先验知识和数据有机地结合。同

时利用主动学习机制，贝叶斯网络可以根据学习进程，主动选择最佳样本进行模型学习，从而有效减少所需评价样本的数量。

下面两节将详细介绍应用主动学习方法，贝叶斯网络和动态贝叶斯网络如何构建客户信用评估模型和客户信用风险行为演化模型。

8.2 基于主动贝叶斯网络的电信客户信用风险评估模型

8.2.1 指标集确定

电信客户信用风险评估指标集采取自顶向下的属性筛选思路。根据双变量统计，剔除那些与目标变量相关度较小的变量，之后，对剩下的变量参与建模。

利用某省移动公司2001年1～5月的44220个客户的缴欠费数据进行统计分析。原始数据集中包含限制消费额度(LIMIT_OWE)字段，该字段是运营商根据经验，对不同客户在每个月限定一个通信费用的上限，如果客户的通信费用超出了这个值，则立刻停机。实际上在一定程度上可以把它看作信用额度的概念，侧面地反映了客户的信用好坏。通过对这个字段取值进行了查看，发现在0～2000这个数据段取值是连续的，而在小于0时，分别去了4个孤立值(－5、－15、－100、－200)，在大于2000时，也分别去了4个孤立值(5060、5160、5860、6060)。因此，为了能够利用该字段分析客户信用好坏的数据特征，把这个字段在7个区间内进行离散化，即小于0元、0～300元、300～700元、700～1000元、1000～1500元、1500～2000元、2000元以上，LIMIT_OWE字段的取值落在这些区间的相应取值为"1"、"2"、"3"、"4"、"5"、"6"、"7"。离散化的字段名为LIMIT_OWE_LEVEL。

离散化目标字段以后，利用了双变量统计方法来对数据进行分析，初步确定与LIMIT_OWE_LEVEL字段相关的属性集。使用双变量统计的方法，通过目标变量不同属性值对于其他属性的影响程度的对比，找出随目标变量属性变化差异较大的属性，剔除差异性不大的属性。因此，在原始数据表中，选取付费方式、性别、年龄段、在网时长等级、欠费状态、地区、预付费、优惠费、应缴费、月租费、特服费、市话费、漫游费、长话费、附加费、农话费、频占费、信息费做双变量统计，图8.2所示为利用IBM DB2 Intelligent Miner中的双变量统计工具所得到的全局分析的结果。

通过对该图的分析可以看出，影响客户消费限额多少的主要特征属性有(16个)性别、年龄、在网时间、付费方式、应缴费、月租费、预付费、市话费、漫游费、长话费、频占费、优惠费、地区、欠费状态、新交费、PAYED_PREPAY。

8.2.2 基于投票熵和类条件后验最大熵的主动贝叶斯网络建立客户信用度评分模型

1. 问题背景

客户信用度是客户履约的可能性，分数越高说明信用越好，违约的可能性越小。在电信企业的经营分析决策支持系统中，客户信用度是用于评价客户欠费风险的指标，它表示客户因各种原因而不如约缴纳话费和拖欠话费可能性的度量。通过对客户的自然属性、消费行为和呼叫行为等进行分析，建立准确的信用度评分数学模型，并将信用度评分模型应用于未

知客户，得到其信用度值，分值范围为0～100。

目前我国电信行业刚刚开始进行信用风险评估工作，大量带有信用风险类标签的历史客户数据并不存在，因此通过传统的分类算法建立信用风险评估模型来预测新客户的信用风险并不可行。为了评价客户的信用风险，只能由领域专家手工评价标注，这是件极其耗时和费力的事情。然而，主动学习可以通过抽样选择未带类标注样本，使用少量的训练数据来建立分类模型，减少了标注样本所付出的代价。

因此，利用6.2.2小节中的基于投票熵和类条件后验最大熵相结合的QBC主动贝叶斯网络分类器，对电信客户信用度建立评分模型，模型输出“有信用”和“无信用”两种分类类别，同时贝叶斯网络计算出属于某一类别的概率，把客户属于“有信用”类别的概率扩大100倍后，得出的分值看作该客户的信用度，例如，某客户通过贝叶斯网络模型预测无信用概率0.1，有信用概率0.9，则该客户的信用度分值为90。如此得到的信用度分值是基于客户历史数据，并且利用贝叶斯网络从中发现其规律而计算得到的，分值客观，防止了个人主观态度的影响。

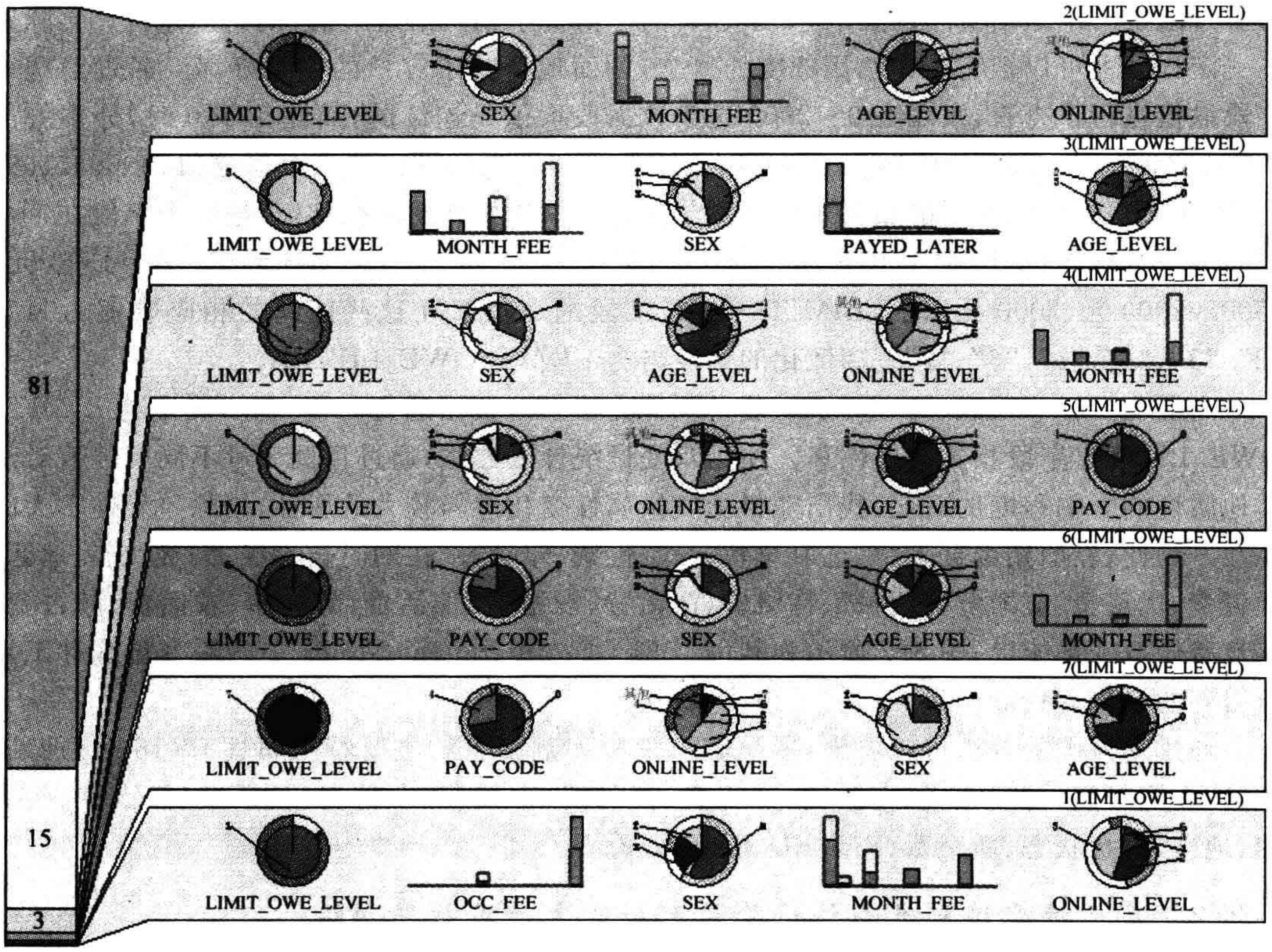


图8.2 全局分析的结果

注：柱状图形中深色对应的部分表示的全体客户的分布情况。柱状图的横轴表示数据层次，纵轴表示对应层次在群体中的比例。饼状图的外圈表示全体客户的分布，内圈表示该群集客户的分布。

2. 算法原理

基于投票熵和类条件后验最大熵相结合的QBC主动贝叶斯网络分类算法具体步骤见6.2.2小节，这里给出其基本流程框图见图8.3。

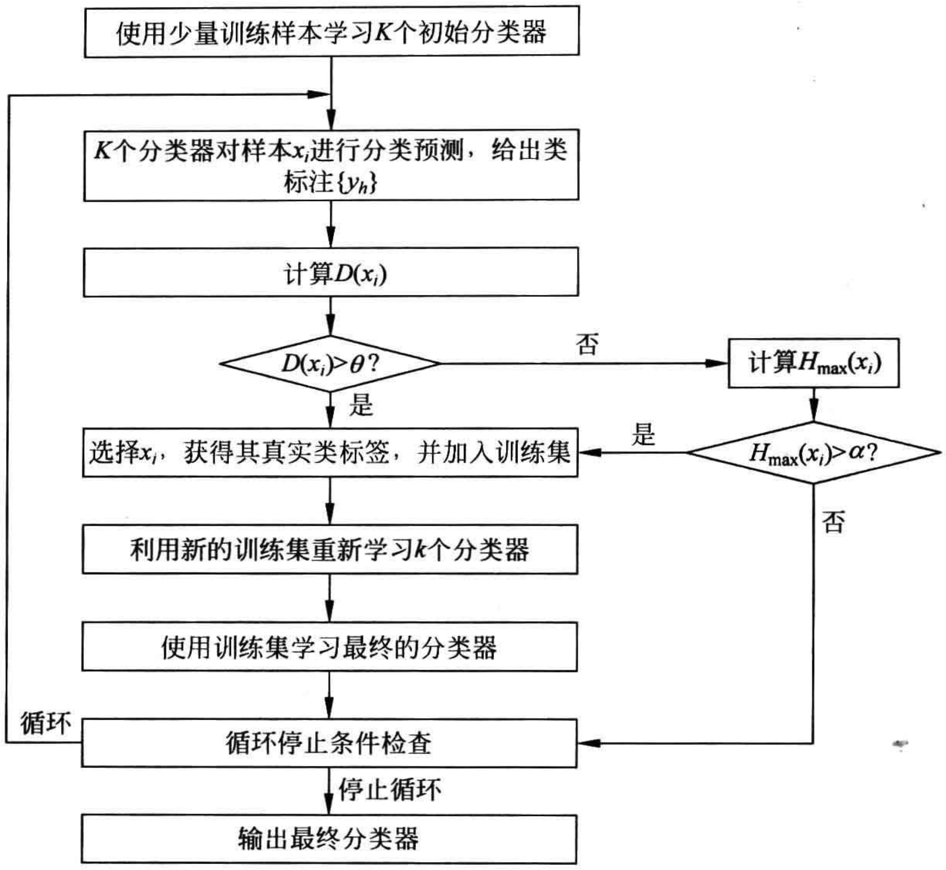


图 8.3 基于投票熵和类条件后验最大熵相结合的 QBC 主动学习算法流程框图

3. 应用结果和分析

实验使用某省移动通信公司 2001 年 1～5 月的 3733 条客户数据进行信用度评分建模。其中 2573 个客户数据作为训练数据(随机抽取 578 条为初始带类标签数据，其余 1995 条为看作未带类标注数据)无信用客户 1030 个，有信用客户 1543 个；1160 个客户数据作为测试数据，无信用客户 464 个，有信用客户 696 个。可用的样本数据属性共有 23 个，使用双变量统计选取的付费方式、性别、年龄、在网时长、欠费状态、地区、月租费、市话费、漫游费、长话费等 16 个字段参与建模，并且对连续字段进行了离散化。

委员会成员通过随机抽取训练数据(Bagging)建立两个分类器来构建。投票熵的阈值 $\theta \geqslant 0.5$，$H_{max}(\boldsymbol{x}_i)$的阈值 $\alpha > 0.6730$，选择停止的条件 ζ 为分类准确性不小于 95%或者搜索完所有的未带类别标注的样本集。

图 8.4 所示为该组数据的 3 种学习算法——基于投票熵的 QBC 方法(VE)、基于投票熵与类条件后验最大熵相结合的 QBC 方法(VE&H_{max})和被动学习的分类准确率曲线比较。从图 8.4 中可以看到，基于投票熵的 QBC 方法选择停止后，总共选择了 71 条未带类别标注的样本，分类预测准确率为 88%；基于投票熵与类条件后验最大熵相结合的 QBC 方法选择停止后，总共选择了 111 个未带类别标注的样本，分类预测准确率达到了 90%。被动学习使用了除测试数据以外的 2573 条记录作为训练数据，分类准确率为 90%。表 8.1 所示为 VE&H_{max}模型在测试集上的各类召回率比较，从表 8.1 中可以看出，该模型对各类的召回率都相当高。

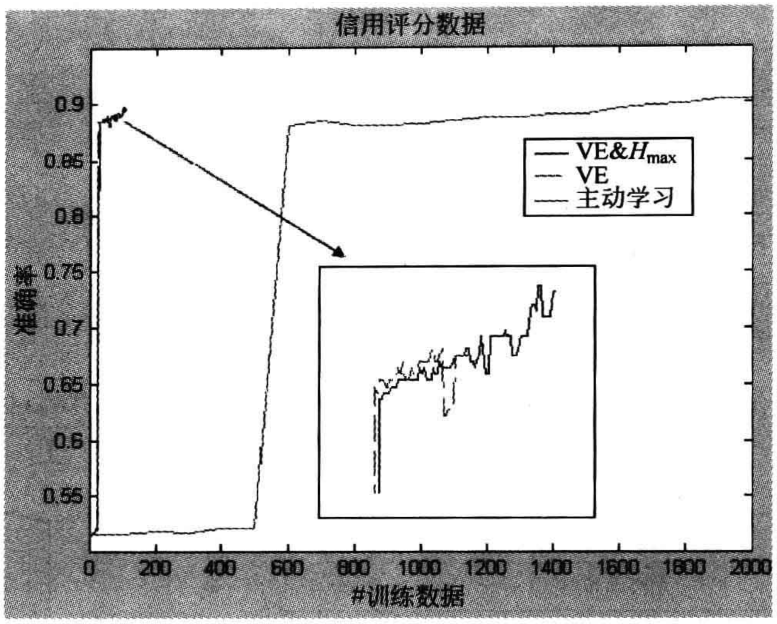


图 8.4 3种学习算法分类准确率曲线比较

表 8.1 VE&H_{max}模型测试结果

信用类别	无信用用户	有信用用户	召回率
无信用用户	399	65	86%
有信用用户	51	645	93%

VE&H_{max}模型结构如图 8.5 所示，图中结点 16 为信用类别结点。图 8.6 所示为 VE&H_{max}模型在测试集(1200 个客户数据，其中无信用客户 138 个，有信用客户 1062 个)上的客户信用度分值分布，从图 8.6 中可以看出，90%的以上计算的信用度落在高端，这与实际情况相吻合，说明 VE&H_{max}算法建立的模型有很高的准确性。图 8.7 所示为带有信用度评分字段的预测数据表。其中字段 16 为客户实际的信用类别字段，字段 17 为 VE&H_{max}模型预测的客户信用类别，字段 18 为 VE&H_{max}模型评估的客户信用度值。

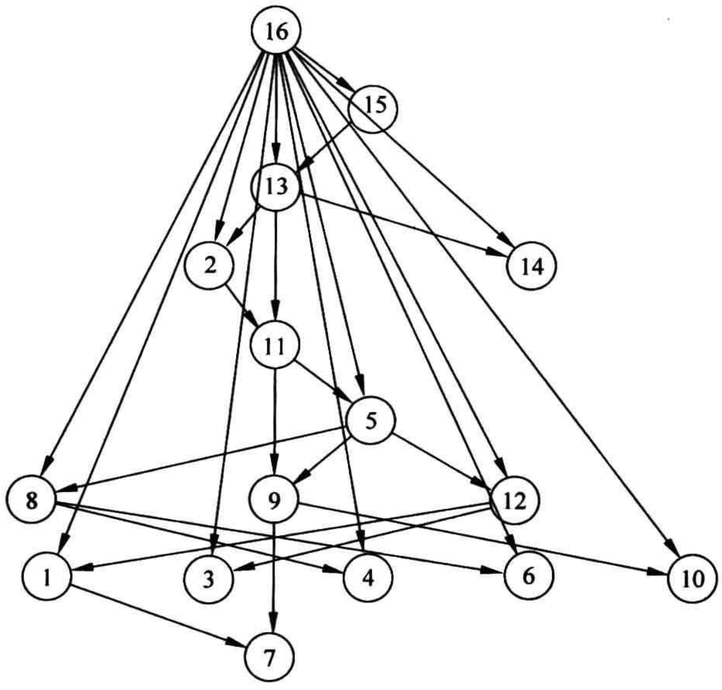


图 8.5 电信客户信用评分模型结构

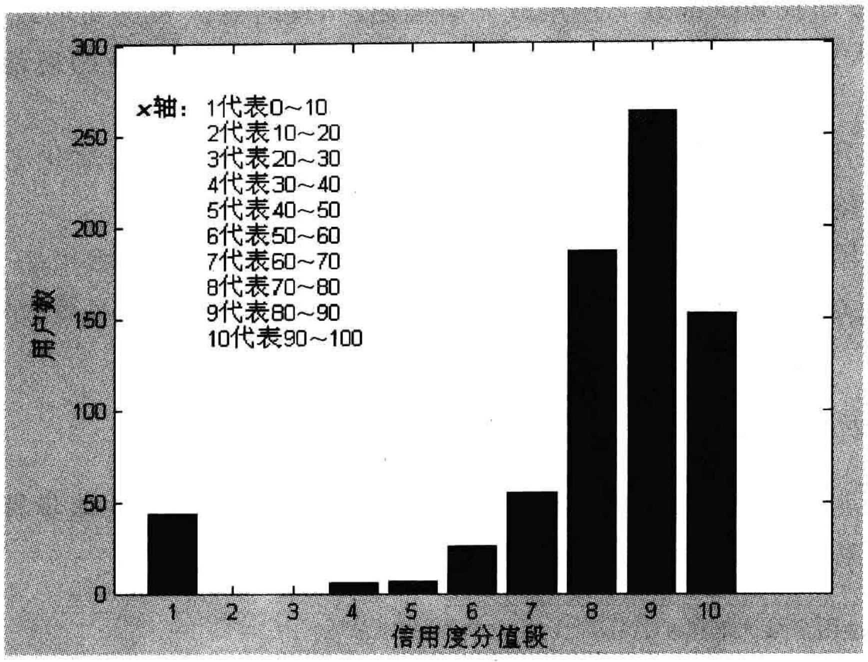


图 8.6 客户信用度分布

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	3	1	2	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	3	1	1	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	3	1	2	2	1	1	2	2	3	1	1	5	2	2	2	50.378
1	1	3	1	2	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	1	2	1	1	1	1	3	1	1	5	2	2	2	82.364
1	1	3	1	2	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	1	2	1	1	1	1	1	1	1	4	2	2	2	90.606
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	5	1	2	2	1	1	1	1	3	1	1	5	2	2	2	82.364
2	2	2	1	2	2	1	2	1	1	3	1	1	4	2	1	1	36.966
1	1	3	2	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	1	2	1	1	1	1	3	1	1	4	2	2	2	77.813
1	1	3	1	1	2	1	1	1	1	1	1	3	2	2	2	2	100
1	1	2	1	2	2	1	1	1	1	3	1	1	3	2	2	2	83.369
1	1	2	1	1	2	1	1	1	1	1	1	2	4	2	2	2	79.945
1	1	3	1	1	2	1	1	1	1	3	1	2	3	2	2	2	73.266
1	1	2	1	1	2	1	1	1	1	1	1	1	5	2	1	2	100
1	1	3	1	2	2	1	2	1	1	3	1	1	3	2	2	2	85.004
1	1	3	1	2	2	1	2	1	1	3	1	1	4	2	2	2	77.08
1	2	3	1	1	2	1	1	1	1	1	1	1	3	2	2	2	91.348
1	1	2	1	2	2	1	1	1	1	3	2	1	3	2	2	2	100
1	1	2	1	1	2	1	1	1	1	1	1	1	4	2	1	2	90.606

图 8.7 信用度评分模型预测结果

8.2.3 基于投票熵和最小相对熵的主动贝叶斯网络建立客户信用等级评估模型

1. 问题背景

客户信用风险等级是电信企业决策支持系统中进行客户总量分析、客户新增与流失分析、客户行为分析、经营风险分析和客户欺诈分析的一个重要分析角度，其可以应用于各种信用风险管理和控制。准确地评估客户信用等级，可以解决在复杂的用户消费群体中良莠

难辨的问题，一方面找到少部分低信用或无信用用户，使恶意欠费透支的现象得到抑制；另一方面，发现绝大多数高信用和按时交费的用户，使他们得到更优质的通信服务，防止他们流失。

建立客户信用风险等级评估模型，主要是为客户信用风险行为演化 DBN 模型中初始网的建立提供网络结构和参数。

客户信用风险等级含有多个等级类别，基于投票熵和最小相对熵相结合的主动贝叶斯网络分类算法更适合于多类分问题，因此，采用该算法作为信用风险等级评估模型的建模算法。

2. *算法原理*

基于投票熵和最小相对熵相结合的主动贝叶斯网络分类算法具体步骤见 6.2.2 节，这里给出其基本流程框图如图 8.8 所示。

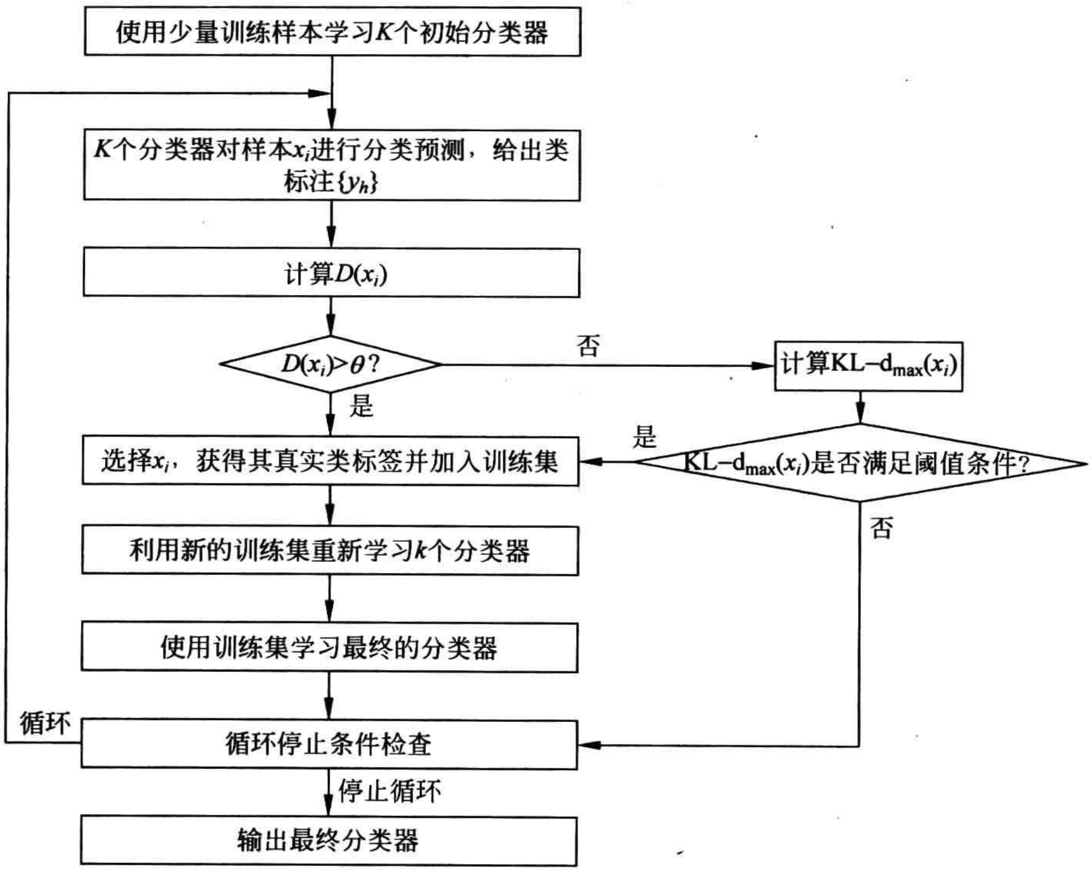


图 8.8 基于投票熵和最小相对熵相结合的主动贝叶斯网络分类算法流程框图

3. *应用结果和分析*

实验使用某省移动通信公司 2001 年 1～5 月的 33512 条客户数据进行信用等级分类的建模。电信专家根据客户的基本情况、缴欠费记录和呼叫行为记录对客户的信用等级进行了评估，信用等级分为 4 个级别(数据类分布见表 8.2)。对其中 4755 条数据作为训练数据，把这个数据集再分为两个部分，其中 3743 条记录看作主动学习过程中的未带类别标注的候选样本集，另外 1012 条记录看作少量带有类标注的样本集，并用这个数据集再学习两

个初始化分类器。测试数据集为28757条客户记录。选取付费方式、性别、年龄、在网时长、欠费状态、地区、月租费、市话费、漫游费、长话费等16个字段参与建模,并且对连续字段进行了离散化。

表8.2 电信客户信用等级数据类分布

类	N	$N\%$
1	22788	68%
2	9717	29%
3	905	2.7%
4	102	0.3%

委员会成员为两个,分类算法采用MBBC贝叶斯网络分类器,阈值$\theta \geqslant 0.5$,$\alpha > -0.0031$和$\alpha < 0.05$,$\zeta \geqslant 85\%$。

实验对基于投票熵的QBC方法(VE)、基于投票熵与最小相对熵相结合的QBC方法(VE&KL-d_{min})和被动学习进行了实验比较,其中基于投票熵的QBC方法选择了476条样本数据,分类精度达到81%,基于投票熵与最小相对熵相结合的QBC方法选择了947条样本数据,分类精度达到84%,被动学习方法使用了所有的3743条样本数据,分类精度达到84%。三者的比较见图8.9。基于投票熵与最小相对熵相结合的QBC方法所建的信用等级评估模型结构如图8.10所示。表8.3所示为VE&KL-d_{min}模型在测试集上的测试结果,其中对等级3和等级4的召回率不是很高,原因是训练数据集中这两类的数据比例与前两者比很小,故这两类数据大部分被分到了等级1和等级2中。为了建立更加准确的模型,需要调整训练数据集中各类的数据比例,但需要长时间的积累后两类的数据。

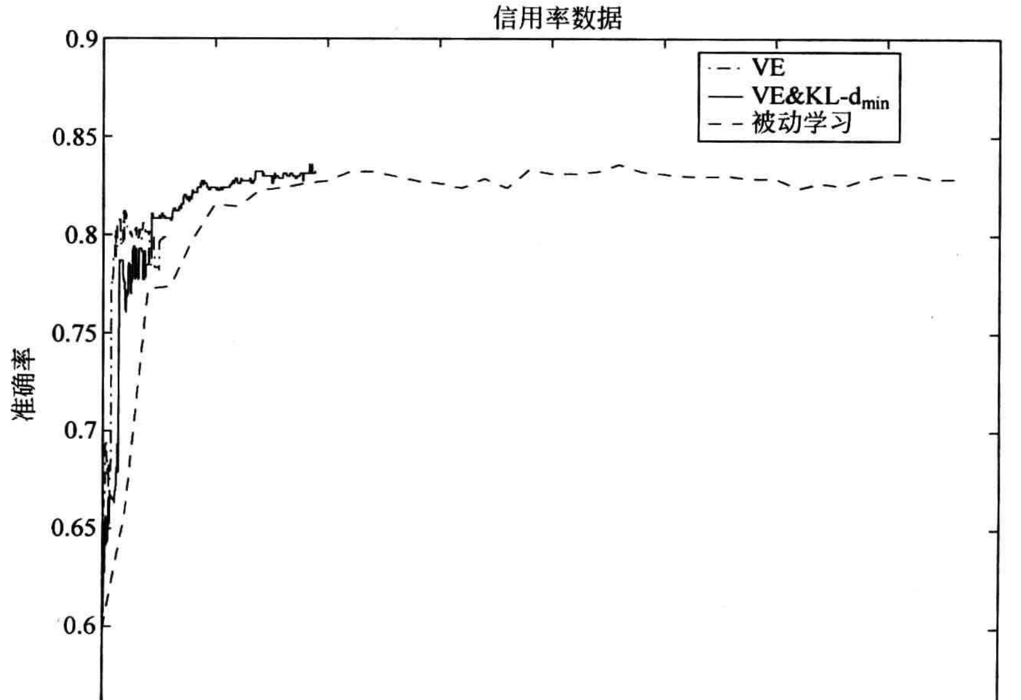


图8.9 3种学习方法的分类准确率比较

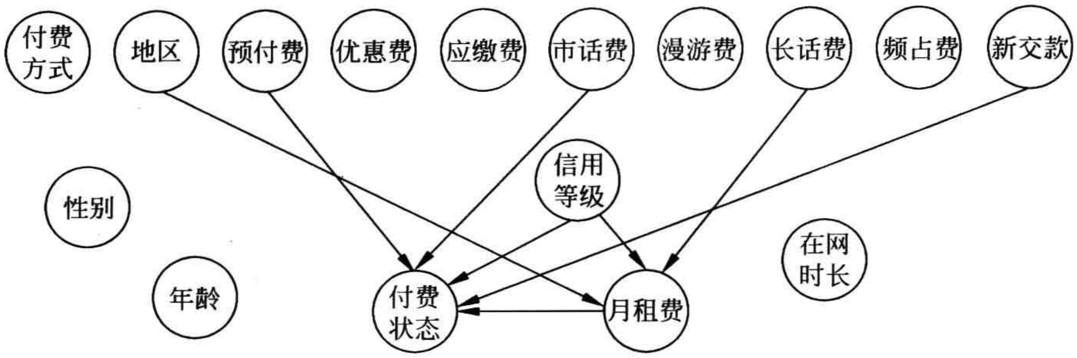


图 8.10 信用等级评估模型结构

表 8.3 VE&KL-d_{min} 模型测试结果

信用等级	1	2	3	4	不能分类	召回率
1	18616	824	52	33	30	95%
2	2668	5256	185	208	22	63%
3	304	243	158	51	21	20%
4	33	15	9	26	3	30%

但是,从总的实验结果来看,对于多类分问题,基于投票熵与最小相对熵相结合的 QBC 方法比原有 QBC 方法更好,能够达到与被动学习相同的分类准确性,并且由主动学习选择的少量样本数据有助于分类器性能的提高,主动学习模型使用的训练数据量远远比被动学习要少,这极大地降低了人为标注样本标签的工作量。

8.3 基于半监督主动学习的 DBNs 建立电信客户信用风险行为演化模型

8.3.1 问题背景

对于电信客户的信用评估问题,在网客户的信用等级是一个动态变化的过程,客户消费行为的变化表现了其信用级别相应的变化情况。客户信用风险行为演化模型就是要通过一定时期的客户行为表现来预测客户在未来各时间点的信用风险趋势。采用动态贝叶斯网络来建立客户行为评分模型,不仅可以找到客户信用等级的变化规律,同时可以对信用风险发生的原因及其结果进行分析,为制定短期或长期的经营管理决策提供依据。

该模型对单个账户信用行为统计分析,对现有客户未来欠账风险的评估,主要考察老客户的信用风险水平的波动性。首先,假定信用等级转换概率是遵循一个稳定的马尔可夫过程(Markov Process)。马尔可夫过程是指客户的信用等级转换至其他信用等级的概率与它们过去的此种概率是没有任何相关性的。

对于信用风险行为演化模型,时间跨度的选择更是关键,因为行为分析是一个纵长的预报系统,假设一个行为模型是基于输出时间段是 12 个月的样本构建的,那么这个模型可能被用来预测在未来 12 个月内,当前的风险级别是否会出现变化。一般行为评分模型的时间段为 6 个月到 2 年。在此,由于实验数据来源有限,选定时间跨度为 5 个月。

由于目前电信客户信用评估刚刚展开,客户的基本数据还不完备,如婚姻状况、学历、薪

金、其他欠贷款情况等数据目前难以获得，这些属性专家认为也能够影响到其信用等级的评估和信用风险的波动。其次，假设信用等级转换概率矩阵是稳定的，也就是说假定不同客户、不同时期之间的信用等级转换概率是不变的。然而实际上，行业因素、国家因素及商业周期因素均会对信用等级转换概率产生重要影响。因此，在建模时，这些变量是不可观测到的，并且能观测到的部分变量的演化并不一定满足马尔可夫假设，这样尽管潜在的客户信用行为模型满足马尔可夫假设，但由于只观测到了部分变量，也无法用这些变量建立准确的动态贝叶斯网络模型。然而，这些因素可以在模型中以隐藏变量的形式出现，因此，有必要增加隐藏变量来参与建模。同时，由于专家评估客户的数量有限，大量客户数据未带有信用等级标注，因此，应利用主动学习机制仔细挑选样本，通过专家评定有利于建模的样本数据，使其加入训练集中，来加快建模过程，减少人工标注代价。

下面介绍利用 6.2.3 节提出的 DBNs-SSAL 算法来建立电信客户信用风险行为演化模型，预测未来某一时期客户的信用风险状态，考察客户的信用风险水平的波动性。8.3.2 节和 8.3.3 节分别讨论了基于分类误差损失减少和 EM 的 DBNs 学习算法以及基于 QBC 和 EM 的 DBNs 学习算法，通过实际的电信客户信用数据比较了两个算法的计算效率和分类性能。

8.3.2 基于 ERS 和 EM 的 DBNs 学习算法建立客户信用风险行为演化模型

1. 算法原理

基于误差减少和 EM 的动态贝叶斯网络学习算法具体步骤见 6.2.3 节，这里给出其基本流程框图见图 8.11。

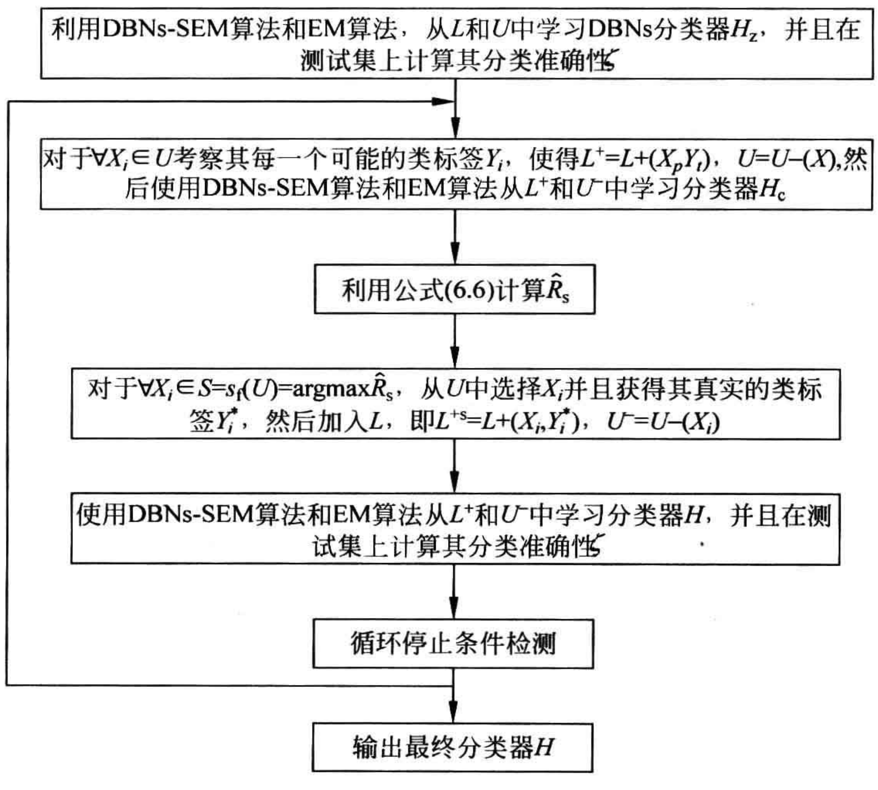


图 8.11 基于误差减少和 EM 的动态贝叶斯网络学习算法的流程框图

2. 应用结果和分析

建模使用某省移动通信公司2001年1～5月的475个客户数据，总共2375条数据记录进行信用风险行为建模。其中1980条数据用作训练数据，在这部分数据中，又随机抽取330条数据记录用作初始分类器的训练，1650条数据看作未标注的数据记录。测试数据为395条数据记录。电信专家根据客户的基本情况、缴欠费记录和呼叫行为记录对客户的信用等级进行了评估，信用等级分为4个级别。选取欠费状态、地区、月租费、市话费、预付费、长话费、新交款和信用级别等8个字段参与建模，并且增加了两个隐结点参与建模。

建模选择8.2.2小节的信用等级评估模型作为初始网的结构，且假设每个时间片内网络结构和参数与初始网相同，整个动态贝叶斯网络按5个时间片展开。

为了验证基于ERS和EM的动态贝叶斯网络学习算法[DBNs-SSAL(ERS&EM)]的效率和性能，实验比较了基于DBNs-SSAL(ERS&EM)算法、基于投票熵的QBC主动学习、基于半监督学习和基于监督学习的动态贝叶斯网络分类器的分类准确性。实验结果如图8.12所示。由于DBNs-SSAL(ERS&EM)的计算复杂度较高，在个人计算机上运行了166h，并未得出最终结果，只选择了7个样本提交用户标记，并将其加入训练集中参与训练。

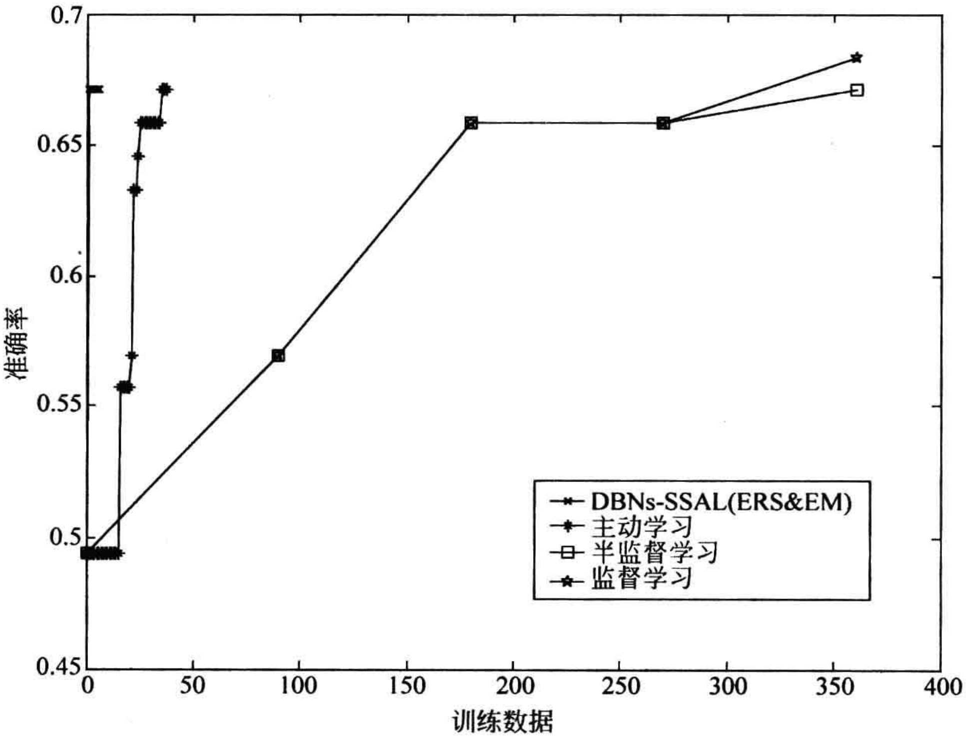


图8.12 4种学习方法的学习结果比较

从实验结果来看，利用半监督学习在学习过程中加入未标记样本时，与加入标记样本的监督学习有相同的性能，但其后续表现不如监督学习，这是因为学习过程中错误分类信息的加入而影响了分类精度。主动学习与半监督和监督学习相比，其学习速度明显快于两者，但是其分类精度不如监督学习。而基于DBNs-SSAL(ERS&EM)的学习算法，其选择了比主动学习更少的未标注样本后，DBNs分类器很快就达到了与其相同的分类精度。因此，从计算效率上来讲，该算法不如基于投票熵的QBC主动学习和基于半监督学习。但是从使用

查询样本的数目上看，比其他 3 种算法要少，并且所获得分类性能好，这无疑减少了人工评价样本的工作量。

8.3.3 基于 QBC 和 EM 的 DBNs 学习算法建立客户信用风险行为演化模型

由于基于 ERS 和 EM 的动态贝叶斯网络学习算法计算复杂度高，建模需要的时间长，因此，本小节介绍另一种学习算法——基于 VE&KL-d_{min}的 QBC 和 EM 的动态贝叶斯网络学习算法[DBNs-SSAL(VE-KLd_{min}&EM)]，建立客户信用风险行为演化模型，讨论其计算效率和模型分类性能。

1. 算法原理

基于 DBNs-SSAL(VE-KLd_{min}&EM)的学习算法具体步骤见 6.2.3 节，这里给出其流程框图见图 8.13。

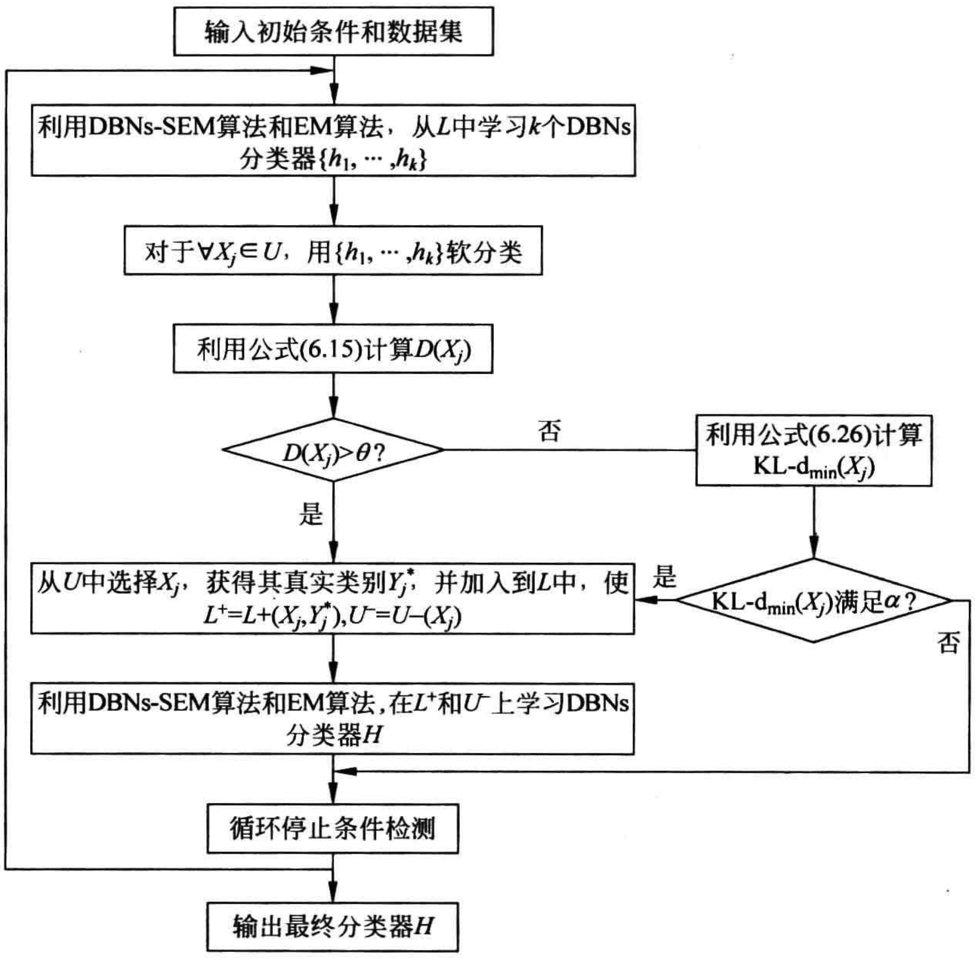


图 8.13 基于 VE&KL-d_{min}的 QBC 和 EM 的动态贝叶斯网络学习算法流程框图

2. 应用结果和分析

实验使用的数据与 8.1.2 节相同，实验比较了基于 DBNs-SSAL(VE-KLd_{min}&EM)算法、基于投票熵的 QBC 主动学习、基于半监督学习和基于监督学习的动态贝叶斯网络分类器的分类准确性。基于 DBNs-SSAL(VE-KLd_{min}&EM)算法使用了 5.46h 计算出最终结果。实验结果如图 8.14 所示。从实验结果来看，基于 DBNs-SSAL(VE-KLd_{min}&EM)算法

选择了比主动学习更少的未标注样本后，DBNs 分类器很快就达到了与其相同的分类精度，并且随着再选择少量样本后其达到了与监督学习相同的分类精度。因此，无论是效率还是性能，均超过了其他 3 种版本的 DBN 分类器。

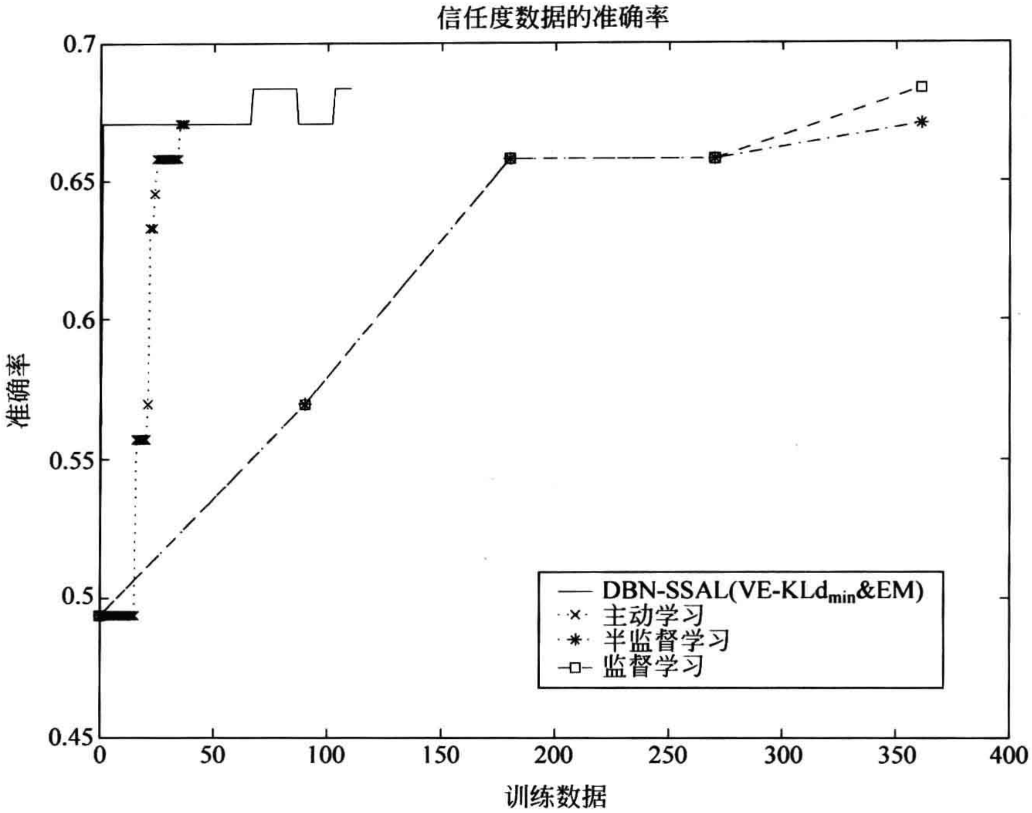


图 8.14 DBNs-SSAL(VE-KLd_{min}&EM)、主动学习、EM 算法和监督学习的学习曲线比较

下面的实验比较了基于 DBNs-SSAL(VE-KLd_{min}&EM)算法和基于 DBNs-SSAL(VE&EM)算法。实验结果如图 8.15 所示，从实验结果来看，VE-KLd_{min}&EM 算法比 VE&EM 算法学习速度快，并且所建分类器的性能更好，可以达到与监督学习相同的分类精度。图 8.16 所示为 DBNs-SSAL(VE-KLd_{min}&EM)算法所建的客户信用风险行为演化 DBNs 模型，其中模型按 5 个时间片展开，h 为信用等级类结点，i、j 为隐藏变量。

同时，实验又比较了基于 DBNs-SSAL(ERS&EM)算法和基于 DBNs-SSAL(VE-KLd_{min}&EM)算法的学习时间和学习效果。从实验结果来看，DBNs-SSAL(ERS&EM)算法的学习时间远远比 DBNs-SSAL(VE-KLd_{min}&EM)算法的时间要长，由于实验条件有限，实验未能得到 DBNs-SSAL(ELRS&EM)算法的最终表现，但从前期的学习效果来看，其与 VE&KL-d_{min}的学习效果相同。

从模型的应用角度来看，如果知道某客户的某个月的信用等级，就可以预测其后 4 个月的信用风险水平。

总之，通过以上的介绍可以看到，基于 QBC 和 EM 的 DBNs 学习算法建立的客户信用风险行为演化模型，从效率和性能上都优于基于误差减少和 EM 的动态贝叶斯网络模型，该建模方法较好地解决了利用监督学习建立信用风险行为演化模型时所面临的训练样本缺乏的困难，解决了人工评价样本所需时间多又代价昂贵的问题，为其实际应用提供了良好的解决思路。

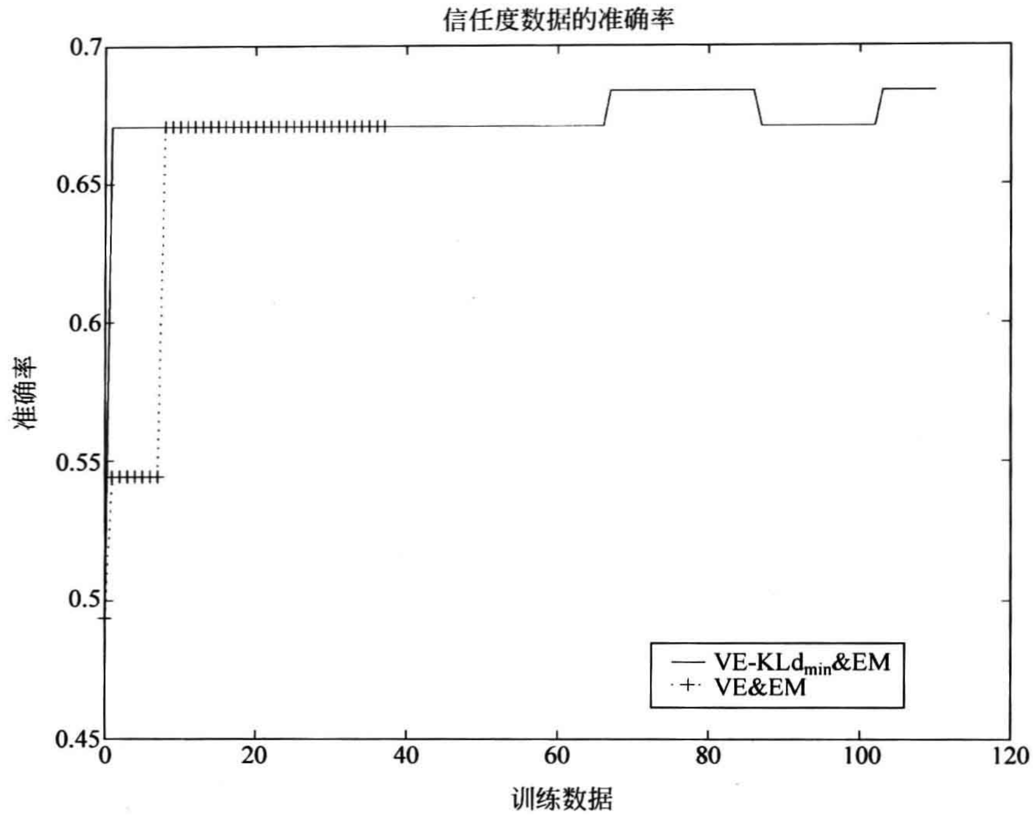


图 8.15 基于 VE&KL-d_{min}与 EM 算法和基于 VE 与 EM 算法的 DBN-SSAL 学习曲线比较

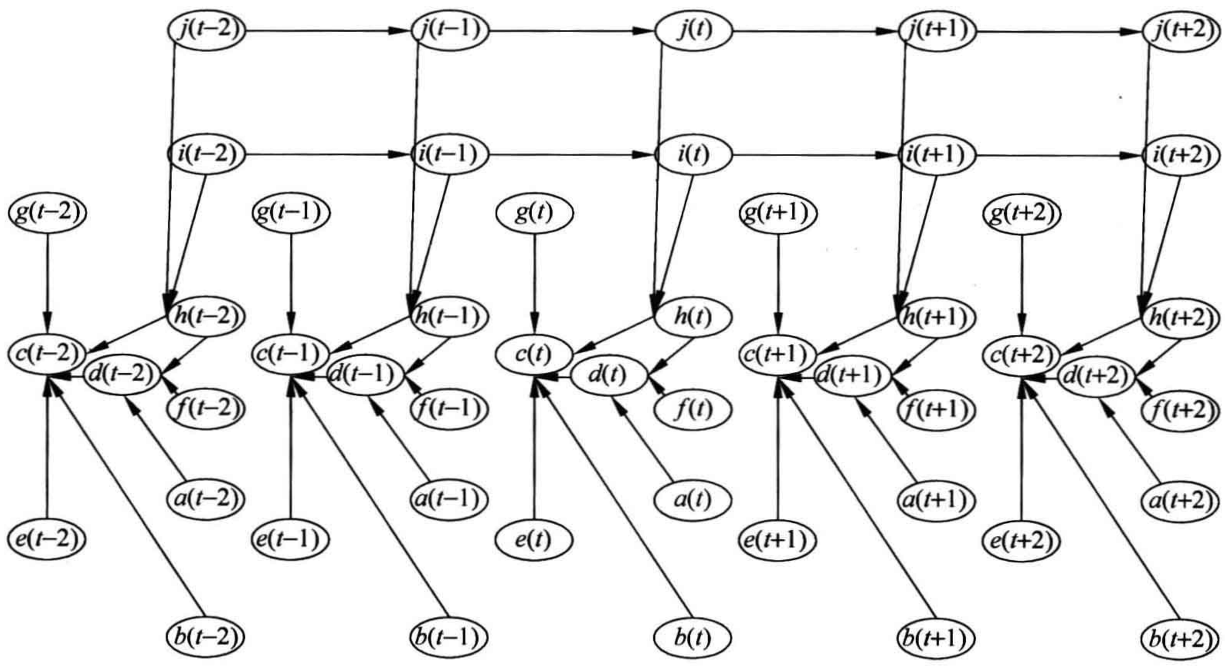


图 8.16 客户信用风险行为演化 DBNs 模型

第9章
概率图模型理论在语音识别中的应用

在语音识别领域，基于隐马尔可夫模型的语音识别已经相当成熟，并获得较高的识别结果，然而，随着实验环境的要求，尤其在噪声环境下，识别率骤然下降，怎样降低噪声的影响、提高语音的识别率，是研究者需要解决的问题。伴随语音的视觉特征，其作为一种辅助语音识别有用的信息，引起了研究者的注意，实验表明，在噪声环境下，视觉信息完全不受噪声的影响，因而利用视觉信息的这一特性，可以弥补语音识别中听觉特征对噪声鲁棒性较差的缺点，有效地提高在噪声环境下语音识别的识别率。

本章首先阐述了语音识别的基本原理，对语音识别的各个步骤进行了简要介绍，然后详细介绍了基于 HMM 的语音识别方法和实验效果，对 HMM 的训练和 HMM 的识别进行了阐述，最后介绍了基于深度动态贝叶斯网络的视听语音识别方法和实验效果，主要对深度动态贝叶斯网络的结构拓扑和学习方法、基于深度动态贝叶斯网络的特征学习和融合识别算法进行了详细介绍，同时针对藏语视听语音识别的应用，分析了其模型优势。

9.1 语音识别基本原理

语音识别系统的基本任务就是将说话者的模拟语音信号，识别成文字符号进行输出，基本流程如图 9.1 所示，粗略地表示了一个语音识别系统的结构框图。

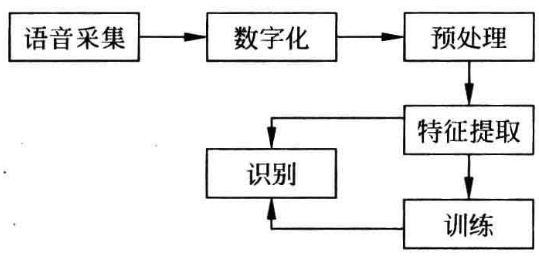


图 9.1　语音识别流程示意图

(1) 语音采集和数字化。说话者发出来的语音信号是模拟信号，而计算机处理的一般是数字信号。语音信号采集就是将输入的模拟语音信号转换成便于计算机处理的数字信号，为系统提供数据，然后将经过处理后的数字信号提交给语音预处理模块。另外，采集模块在获得语音信号之后，还需要对信号进行简单的处理，如滤波、语音格式、标准化等。

(2) 预处理。一般包括预加重、加窗分帧、端点检测。经预处理后的信号被转换成了帧序列的加窗的短时信号。

(3) 特征提取。这是对短时信号进行分析，提取语音特征参数的过程，典型的语音特征参数包括语音帧能量、基音频率、语音信号的短时谱、线性预测系数(LPC)、共振峰频率及带宽、LPC 倒谱特征(LPCC)、MEL 倒谱特征(MFCC)、LSP 特征、基音轮廓特征、部分相关系数特征等。

(4) 训练是在使用过程中不断修订特征提取参数，以使各项参数更加合理，从而提高识别率。

(5) 语音识别是目标语音根据特征参数与模型库中的参数进行匹配，产生识别结果的过程。一般有模板匹配法、概率图模型法和神经网络等。

9.2 基于隐马尔可夫模型的孤立词语音识别

隐马尔可夫模型(Hidden Markov Models，HMM)作为语音信号的一种统计模型，今天正在语音处理各个领域中获得广泛的应用。

对于特定人孤立词识别，DTW 和 HMM 在相同的环境条件下，识别效果相差不大。但对于非特定人孤立词识别，HMM 一方面用隐含的状态对应于声学层各相对稳定的发音单位，并通过状态转移和状态驻留来描述发音的变化；另一方面它引入了概率统计模型，不再用 DTW 算法中动态时间对齐的方法求匹配距离，而是用概率密度函数计算语音参数对 HMM 模型的输出概率，通过搜索最佳状态序列，以最大后验概率为准则找到识别结果，具有更好的性能。

HMM 的缺点在于统计模型的建立需要依赖一个较大的语音库，这在实际工作中占有很大的工作量，且模型所需要的存储量和匹配计算(包括特征矢量的输出概率计算)的运算量相对较大。

在 2.2.1 节中将隐马尔可夫模型分为离散 HMM (Discrete HMM，DHMM)和连续 HMM(Continuous HMM，CHMM)。CHMM 通常是连续高斯混合密度 HMM 模型，其观测矢量服从连续高斯混合密度函数分布，通过选取几个中心不同的高斯混合密度函数，即用多个高斯加权和来近似观测矢量的概率分布，增强了 HMM 的建模能力。因此，本节使用 CHMM 模型进行语音识别。

9.2.1 基于 HMM 的语音识别框架

如图 9.2 所示，利用 HMM 方法进行语音识别，分为两个阶段进行。在训练阶段，将输入的语音信号样本通过预处理、特征参数提取等一定的处理之后，使用 HMM 训练算法，为每个词条建立一个 HMM 模型。当所有词条训练完毕后，将模型数据保存，分别建立了系统词汇表中每个词对应的 HMM 模型，构成了模型库。在识别阶段，将未知语音信号通过同样的预处理和特征参数提取之后，使用 HMM 识别算法，与训练阶段构成的模型库中的模型进行匹配，根据一定的规则，将语音的识别结果输出。

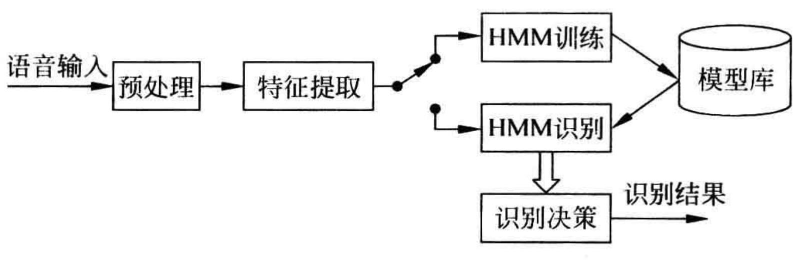


图 9.2 基于 HMM 的语音识别框架

当分析出语音识别结果之后，将获得的识别结果传递给结果处理模块，结果处理模块将做出某种反应。比如，以文字形式显示识别结果，或以音频形式播放语音等。

9.2.2 HMM训练

在语音识别中，观测序列就是通过计算得到的一帧帧的语音参数，如MFCC参数。而状态则是在训练阶段事先规定好的不同语音单元。语音单元可以是一个完整的音节，也可以是更为精细的音素。

一个HMM模型由若干个状态组成，随着时间的变化，各个状态之间可以发生转移，也可以在一个状态内驻留。每个观察向量对不同的状态都有相应的输出概率。如图9.3所示的HMM，包含有4个状态 S_1, S_2, S_3, S_4。状态之间或状态自身的转移概率用 a_{ij} 表示，输入观察序列为 $O=O_1, O_2, \cdots, O_T$。每个观察序列是一帧MFCC参数。在这个模型中，序列 $O_1, O_2, \cdots, O_T$ 是可观测的输入序列，称为观察序列，而每一时刻所处的状态却是隐含的。

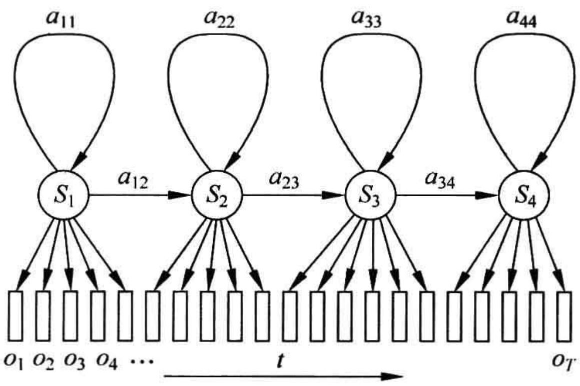


图9.3 HMM与语音参数的关系

HMM模型通常采用 $\lambda=(A,B,\pi)$ 表示，对于一个CHMM，用 N 表示模型的状态数，$\mathbf{A}=\{a_{ij}\}$ 代表状态转移概率矩阵，$\pi=\{\pi_i\}$ 代表各状态的起始概率分布，$B=\{b_j(o)\}$ 代表输出概率密度函数，其中

$$b_j(o) = \sum_{m=1}^{M} c_{jm}\psi(o, \boldsymbol{\mu}_{jm}, \boldsymbol{U}_{jm}) \quad 1 \leqslant j \leqslant N \tag{9.1}$$

式中，M 为每个状态包含的高斯元的个数；c_{jm} 为第 j 个状态第 m 个混合高斯函数的权；ψ 为正态高斯概率密度函数；$\boldsymbol{\mu}_{jm}$ 为第 j 个状态第 m 个混合高斯元的均值矢量；$\boldsymbol{U}_{jm}$ 为第 j 个状态第 m 个混合高斯元的协方差矩阵。

权系数 c_{jm} 满足下面的条件，即

$$\sum_{m}^{M} c_{jm} = 1 \quad 1 \leqslant j \leqslant N \tag{9.2}$$

对于CHMM的每一个状态，都用若干个正态高斯概率密度函数（简称为PDF）的线性组合来表示，每个PDF有各自的均值矢量和协方差矩阵，这些都是通过对大量的MFCC参数进行统计得到的。

HMM训练就是根据输入语音的观察值序列 $O=O_1, O_2, \cdots, O_T$ 和初始模型参数 $\lambda=(\pi, \mathbf{A}, B)$，以一定的规则反复调整模型参数，逐步构成新的模型 $\bar{\lambda}$，使得 $P(O|\bar{\lambda})>P(O|\lambda)$，直至 $P(O|\bar{\lambda})$ 收敛，使概率 $P(O|\lambda)$ 尽可能达到最大，最终得到一个优化模型 $\lambda=(\pi, \mathbf{A}, B)$ 的过程。HMM训练过程可使用前面2.2.1小节中的Baum-Welch算法。但是，模型中的状态随着时间而递增，t 时刻模型处于某一个状态，$t+1$ 时刻模型却处于另一个递增的状态。

然而左右型 HMM 的一个主要问题是不能使用一个单个的观测样本序列来训练模型，换言之，单观测样本序列得到的状态转移统计信息很不充分，不利于模型参数的重估，这是因为模型内部状态的暂态本质仅仅允许较少的观测值用于每一个状态。而 Baum-Welch 算法是在假设只有一个观察序列的条件下推导出来的，因此，为了提高识别精度，采用多样本值，即多观察值序列的 CHMM 方法来构建语音识别系统。

由式(2.9)和式(9.1)可知，CHMM 模型中包含 5 个基本参数：初始状态概率 π_i、状态转移概率矩阵 $\boldsymbol{a}_{ij}$、混合高斯函数权值 c_{jm}，均值$\boldsymbol{\mu}_{jm}$与协方差矩阵 $\boldsymbol{U}_{jm}$。设有 L 个训练样本，通过端点检测、特征提取之后，得到 L 个观察值序列集合，记为 $O=\{O^{(1)},O^{(2)},\cdots,O^{(L)}\}$。其中 $O^{(l)}(1\leqslant l\leqslant L)$表示第 l 个观察样本，长度为 T。那么，由 L 个观察值序列构成的集合表示为 $O^{(l)}=\{O_1^{(l)},O_2^{(l)},\cdots,O_T^{(l)}\}(l=1,2,\cdots,L)$。假定各个观察值序列之间互相独立，则所有观察值序列的联合概率 $P(O|\lambda)$等于各样本输出概率 $P(O^{(l)}|\lambda)$的乘积，即

$$P(O\mid\lambda)=\prod_{l=1}^{L}P(O^{(l)}\mid\lambda) \tag{9.3}$$

那么，多观察值序列的 CHMM 训练，就是求取最佳 π_i、$\boldsymbol{a}_{ij}$、c_{jm}、$\boldsymbol{\mu}_{jm}$ 和 $\boldsymbol{U}_{jm}$ 等参数，根据语音训练样本 $O=\{O^{(1)},O^{(2)},\cdots,O^{(L)}\}$和初始模型参数 $\lambda=(\boldsymbol{A},B,\pi)$，由重估公式求得一组新的 $\bar{\lambda}$，然后判断 $P(O|\bar{\lambda})$是否收敛，如果不收敛，则将新的参数 $\bar{\lambda}$ 作为输入再次进行参数重估，反复修正模型，直到 $P(O|\bar{\lambda})$收敛为止，使这些观察值序列的联合概率达到最大。

定义第 l 个训练样本的过渡概率 $\xi_t^{(l)}(i,j)$为已知观察值序列 $O=\{O^{(1)},O^{(2)},\cdots,O^{(L)}\}$和初始模型参数 $\lambda=(\boldsymbol{A},B,\pi)$条件下，$t$ 时刻处于状态 S_i 和 $t+1$ 时刻处理状态 S_j 时的概率，即

$$\xi_t^{(l)}(i,j)=P(q_t=S_i,q_{t+1}=S_j\mid O^{(l)},\lambda) \tag{9.4}$$

那么根据前向算法递归式(2.17)～式(2.19)和后向算法递归式(2.22)～式(2.24)，得到第 l 个训练样本的前向概率变量 $\alpha_t^{(l)}(i)$和后向概率变量 $\beta_t^{(l)}(i)$，可以计算出过渡概率为

$$\xi_t^{(l)}(i,j)=\frac{\alpha_t^{(l)}(i)a_{ij}b_j(O_{t+1}^{(l)})\beta_{t+1}^{(l)}(j)}{P(O^{(l)}\mid\lambda)} \tag{9.5}$$

由上述公式可以推出，第 l 个训练样本在第 t 时刻处于状态 S_i 的概率 $\xi_t^{(l)}(i)$为

$$\begin{aligned}\xi_t^{(l)}(i)&=P(q_t=S_i/O^{(l)},\lambda)\\&=\sum_{j=1}^{N}\xi_t^{(l)}(i,j)\\&=\frac{\alpha_t^{(l)}(i)\beta_t^{(l)}(i)}{P(O^{(l)}/\lambda)}\end{aligned} \tag{9.6}$$

由此可以得到基于 L 个观察值的重估公式为

$$\bar{\pi}_i=\sum_{l=1}^{L}\xi_t^{(l)}(i) \tag{9.7}$$

$$\bar{a}_{ij}=\frac{\sum_{l=1}^{L}\sum_{t=1}^{T-1}\xi_t^{(l)}(i,j)}{\sum_{l=1}^{L}\sum_{t=1}^{T}\xi_t^{(l)}(i)} \tag{9.8}$$

$$\bar{\mu}_{jm}=\frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\xi_t^{(l)}(j,m)O_t^{(l)}}{\sum_{l=1}^{L}\sum_{t=1}^{T}\xi_t^{(l)}(j,m)} \tag{9.9}$$

$$\bar{c}_{jm} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\xi_t^{(l)}(j,m)}{\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{m=1}^{M}\xi_t^{(l)}(j,m)} \tag{9.10}$$

$$\bar{U}_{jm} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\xi_t^{(l)}(j,m)(O_t^{(l)}-\mu_{jm})(O_t^{(l)}-\mu_{jm})^T}{\sum_{l=1}^{L}\sum_{t=1}^{T}\xi_t^{(l)}(j,m)} \tag{9.11}$$

9.2.3 HMM 识别

需要存入系统词汇表里的每个词条都训练完成之后，记为 $\lambda_1,\lambda_2,\cdots,\lambda_K$，$K$ 为模型库中的模型数目，分别将这些模型参数存储，构成了一个模型库系统。此后，系统就具有了识别词汇表里词条的功能。

HMM 识别的基本任务就是将输入的未知语音信号进行识别。首先通过端点检测获得正确的语音段，然后经过与训练过程中同样的特征提取，得到观察序列 $O=O_1,O_2,\cdots,O_T$，使用前向-后向算法或 Viterbi 算法，计算出观察值序列在训练好的各 HMM 模型 $\lambda_i(1\leqslant i\leqslant K)$ 条件下的输出概率 $P(O|\lambda_i)$，取概率最大的模型为识别词条。

9.2.4 实验结果分析

我们从藏语孤立词视听语音数据集中提取了音频数据，将其用于 CHMM 模型的语音识别评估实验。这个数据集包括 30 个藏文字母、0～9 数字和 10 个命令词 3 个主题。其中，数字语音包含 5 个男生发音，每人对每个数字发音 4 遍；字母语音包含 5 男 5 女的发音，同样也是每人每个字母发音 4 遍；命令词语音包含 10 男 10 女，每人每个词条发音 3 遍。字母语音和数字语音都是噪声语音数据，而命令词语音是纯净语音数据。

所有的语音数据在进行训练之前，都被降采样到 8kHz，每个语音帧帧长为 20ms，帧移 10ms，窗函数采用 Hamming 窗。对于语音特征，采用了 24 维的 MFCC 特征参数。CHMM 的状态数为 3，每个状态的高斯元个数也是 3。

表 9.1 出示了 CHMM 同 DTW 语音识别方法的比较结果。可以看出其具有明显的识别优势。

表 9.1 藏语语音识别结果

方法	Digits	Letters	Words
CHMM	51%	53%	89.5%
DTW	39%	47%	70.5%

9.3 基于深度动态贝叶斯网络无监督语音特征学习和语音识别模型建立

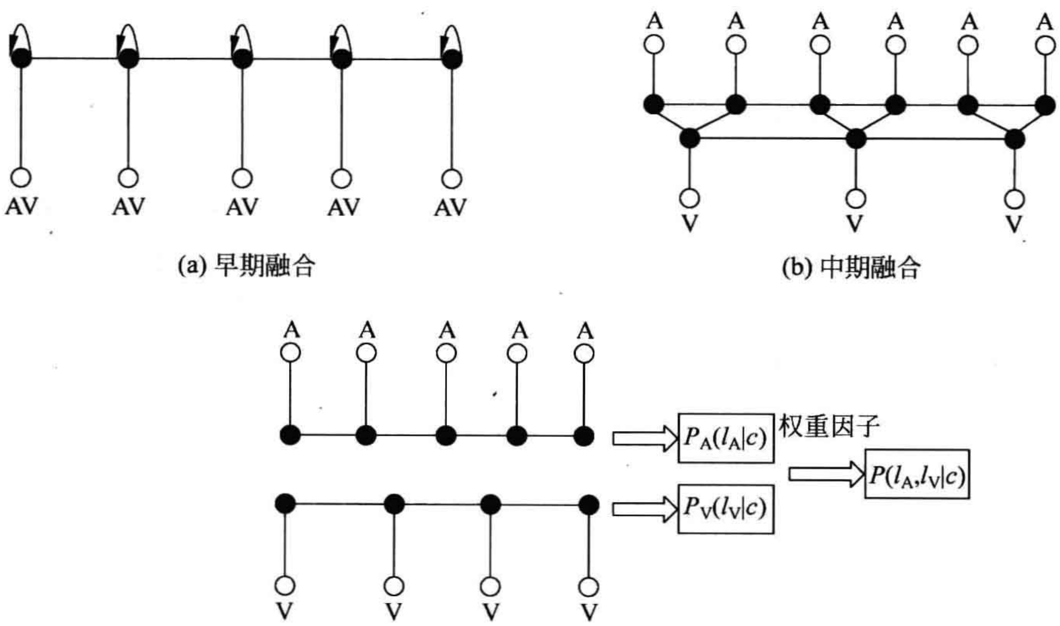
在视听语音识别中，多流动态贝叶斯网络(Multi-stream Dynamic Bayesian Network)和耦合隐马尔可夫模型(Coupled HMM)是常用的方法，但是它们有两个明显的不足：一是

不能学习多模态共享属性，即无法学习多模态数据之间的相关性，也就无法实现跨模态识别和数据重构；二是这两种方法在建模时忽略了帧和帧之间的观测属性的依赖关系，然而，有研究表明，对于语音信号和视频信号，帧间属性是有上下文依赖关系的。针对这些问题，本节介绍的深度动态贝叶斯网络不仅可以无监督学习各模态特征属性和多模态共享属性，而且其不受帧间独立的假设约束，与多流动态贝叶斯网络和 Coupld HMM 模型相比更加适合视听多模态语音识别。

9.3.1　概述

近年来，由于视听语音识别的较强抗噪性特点，其已经成为提高语音识别系统鲁棒性的一条新途径。

目前视听语音识别的多模态信息融合策略分为早、中、晚期融合，如图 9.4 所示。很显然，早、晚期融合没有反映出语音与视觉特征之间的相关性。中期融合是在各模态特征数据的基础上，通过进一步统计计算，学习各模态特征之间的相互关系再进行识别。中期融合有利于人们认识各模态数据之间的相关性，其不仅可以实现多模态融合识别，也可以实现跨模态识别与数据重构，即当某个模态数据（如音频数据）丢失时，利用其他模态数据（如视频数据）和特征之间的相关性，恢复其数据原貌，并给出其类别。这对于现实应用中的唇读识别、音视频数据恢复、语音动画合成都具有十分重要的意义。



(c) 晚期融合

图 9.4　视听语音识别的融合策略

A—音频数据；V—视频数据

现有的中期融合方法大多采用了浅层模型（只有一个隐层的特征向量表示），如 Coupled HMMs 和多流 DBNs，这些模型一是无法学习多模态数据之间的强相关性；二是在建模时这些模型假设帧间特征属性独立。由于视、音频信息的相关性是高度非线性的，并且帧间特征属性实际是由上下文依赖关系，因此这些浅层结构模型难以学习各模态特征之间的复杂关联。

最近几年,深度学习方法被越来越多地应用在复杂非线性的属性学习中,获得了良好的识别效果。其大部分的方法是利用限制玻耳兹曼机(Restricted Boltzmann Machines, RBM)来构建深度信任网络(Deep Belief Networks),从大量未标注的视听语音数据中学习各模态特征之间的深层次关系。虽然实验显示其噪声环境下识别准确率较现有的融合方法提高了近10%,但是它仍存在以下问题:

(1) 空间结构人为设定,不是通过数据驱动学习而得。如果对所研究的语言(本项目背景语言为藏语)发音特征(如音素)和唇部视频特征(如视素)无先验知识时,这将导致学习到的融合特征向量存在冗余或者缺失,以及某些变量之间的关系无法学习。

(2) 不能学习变量的时间变化模式。然而,语音和视频最基本的问题就是时间序列的表示。

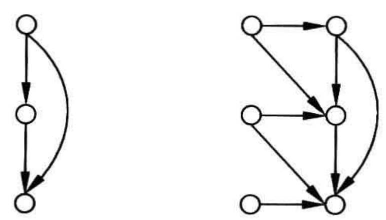
(3) 回调修正学习需要大量带类标注数据。对于少数民族语言,由于缺乏训练有素的语音标注专家,语音标注代价极高,大量带类标注数据难以获得。

因此,针对以上问题,下面介绍一种深度动态贝叶斯网络(Deep Dynamic Bayesian Network, DDBNs),其借助于DBNs强大的时空表达能力及优化的网络结构学习能力,通过与深度学习相结合,可以有效地学习并表示出多模态信息之间的复杂时空相关性。同时该模型结构不受帧间独立假设的制约,可以很好地表现语音特征、视频特征和多模态共享特征的动态变化规律。

9.3.2 深度动态贝叶斯网络

根据2.2.3小节介绍的动态贝叶斯网络的基本原理,知道其拓扑结构分成两部分,如图9.5所示,一部分是空间网络(又叫做初始网络),另一部分是转移网络。

根据DBN的一般拓扑结构,为了实现在不同层次上捕捉变量之间复杂的时空关系,提出了如图9.6所示的深度动态贝叶斯网络。在这个网络中,复杂模式被自底向上分解成不同抽象级别变量之间的时空关系,低层变量的输出是上层变量的输入,上层的变量输出总结了它的输入变量之间的时空链接关系。



(a) 空间网络(初始网络) (b) 转移网络

图9.5 动态贝叶斯网络的基本拓扑结构

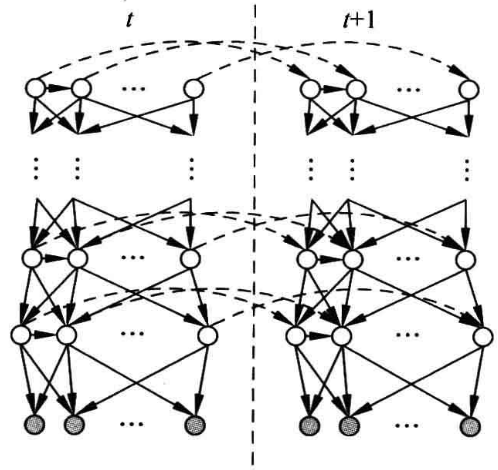


图9.6 深度动态贝叶斯网络拓扑

深度动态贝叶斯网络的层内结点与层间结点之间都可以通过学习建立有向链接关系。最底一层是观测数据输入层,其可以是信号的原始数据或是信号的观测特征数据(如语音原

始数据或语音特征数据、图像原始数据或图像特征数据等)，这层变量是可观测的。从第二层开始的各个层为隐层，隐层是目标模式在不同级别的抽象表示，相邻两层的低层输出作为高层的输入。隐层结点的输出表征了其与链接结点的时空关系。数据输入层再经过上层变量逐级的抽象学习，直到最后一层，可以使用少量的特征变量来表示复杂的模式关系。

基于深度动态贝叶斯网络，又提出了用于视听语音识别的双模态深度动态贝叶斯网络，如图 9.7 所示。其中，最顶一层为输出层，即识别的语音类别，与其紧邻的下一层是共享(融合)特征向量，其与输出层之间形成分类映射关系。模型的最底层为数据输入层，分别来自音频通道和视频通道。考虑到帧间特征变量的依赖关系，假设除数据输入层以外的其他隐层变量遵循马尔可夫第一定律。

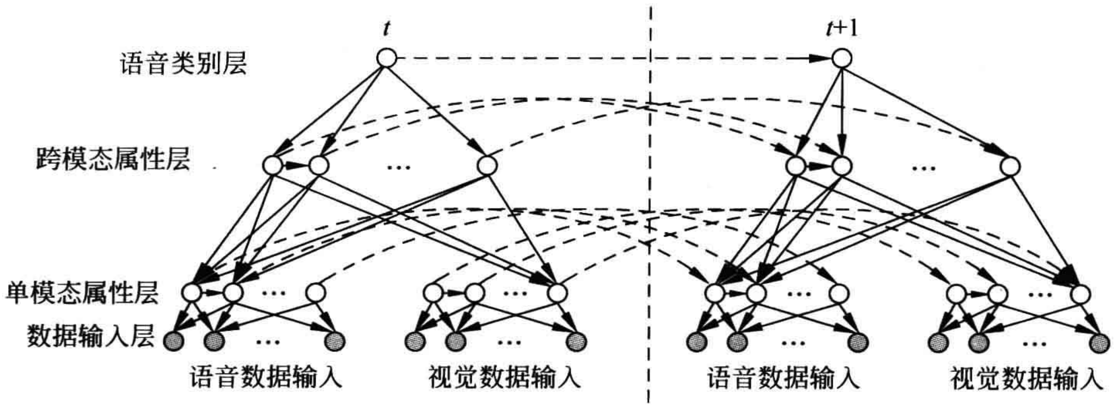


图 9.7 基于深度动态贝叶斯网络的视听语音识别模型

假设低维特征向量嵌入在高维原始数据中，这里假设每一个隐层变量都是一个 l 维的隐连续状态变量。在图 9.7 中，假设 x_i^t 是时间片 t 中数据输入层的第 i 个结点，$h_k^{w,t}$ 是时间片 t 中第 w 层的第 k 个隐变量，s^t 代表最顶层的离散类结点集，根据动态贝叶斯网络的基本原理可以写出对于类结点的概率分布计算公式为

$$P(s^t = z) = \sum_{i=1}^{I} P(s^t = z \mid s^{t-1} = i) p(s^{t-1} = i) \tag{9.12}$$

同时也可以写出在变量集合 $\{x_i^t, h_k^{w,t}, s^t\}$ 上的联合概率分布计算公式为

$$P(\{x_i^t, h_k^{w,t}, s^t\}) =$$

$$\left[P(s^1)\prod_{t=2}^{T} P(s^t \mid s^{t-1})\right]\left[\prod_{t=1}^{T}\prod_{w=1}^{2}\prod_{k=1}^{K} P(h_k^{w,t} \mid \mathrm{Pa}(h_k^{w,t}))\right]\left[\prod_{t=1}^{T}\prod_{i=1}^{I} P(x_i^t \mid \mathrm{Pa}(x_i^t))\right] \tag{9.13}$$

式中，Pa(*)为结点 * 的父变量集。

基于深度 DDBN 的视听语音模型是一个层次化的结构，其在不同抽象层捕捉特征变量之间的时空关系，其允许多模态信息在中期进行融合，并且允许帧间特征向量具有依赖关系，这种结构形式更适合于表示视听语音信号。

9.3.3 DDBN 拓扑结构的学习和视听语音识别学习算法

基于 4.3 节中的动态贝叶斯网络的结构学习方法，提出了无监督深度动态贝叶斯网络的学习方法，使用这个学习方法不仅可以学习各单模态特征属性，也能够无监督学习共享特征属性，同时学习视听语音识别模型。首先介绍 DDBN 的学习方法，然后给出基于 DDBN

的视听语音识别模型学习算法。DDBN 的学习分为两个步骤，即空间网络的学习和转移网络的学习。

1. 空间网络的学习

对于空间网络的学习，按照图 9.8 所示的方式，一个块一个块地进行空间拓扑结构的学习。在每一个块中，需要学习输出层结点的个数和输入层与输出层之间的链接关系。

(1) 特征属性学习

在一个块中，给定底层输入，首先需要决定上层(输出层)结点的个数。最直接的方法是采用贪婪学习方法，从只有一个输出结点开始，使用结构学习算法(2)，学习优化结构，然后每次迭代再增加一个结点，计算相应的优化结构，这个迭代过程直到估计的优化结构的评分不再提高为止。

(2) 结构学习

给定每层结点数目以后，需要学习结点之间的链接关系。假设图 9.9 所示为 DDBNs 的任意一个块结构，其中 A_s 为隐结点，B_s 为输入结点，为了学习结点之间的链接关系，应用 DBNs 结构 EM 学习算法来学习块结构。结构 EM 学习包含两个步骤，即 E-step 和 M-step。E-step 为给定当前网络结构，计算结点 A_s 的期望值，而 M-step 应用计算出的期望值学习一个新的网络结构。这两个步骤迭代循环，直到当前学习到的网络结构评分不再提高，并给出最后的网络结构和参数。

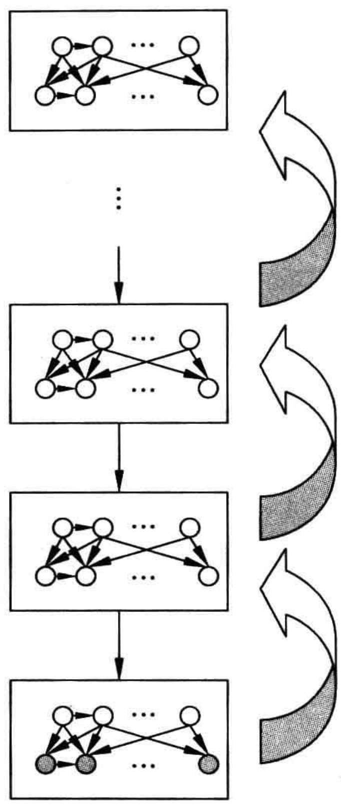
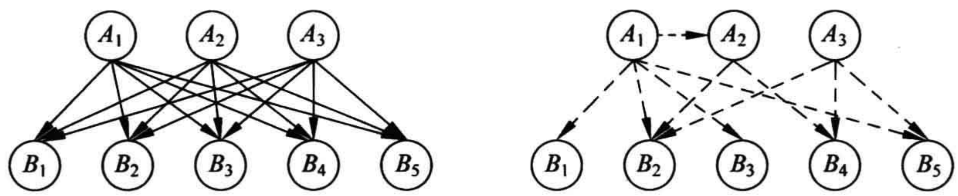


图 9.8 空间网络学习



(a) 初始结构　　(b) 目标结构

图 9.9 块结构学习

下面是结构 EM 算法的总结：

Algorithm9.1 The structural EM algorithm

1. Choose an initial structure G_0 and estimate the parameters θ_0 for G_0。
2. for $t = 0, 1, \cdots$ until convergence
   ① "Complete the data" — Compute the expected values of the variables with structure G_t and parameter θ_t。
   ② Find the structure G_{t+1} that maximize the score (BIC、BDe 等) of the two-layer network through a greedy hill climbing procedure。
   ③ Learn the parameter θ_{t+1} for G_{t+1} with the parameter EM algorithm。

一旦一个块结构学习完成后，就可以对输出层结点进行抽样，产生新的数据作为下一个块的输入层数据，然后不断重复块结构的学习过程，直到最后一个块学习完成。

(3) 视听语音 DDBN 模型的空间结构学习

对于视听语音 DDBN 模型，首先学习两个单模态块的结构，即语音模态块和视频模态块，然后再向上继续学习共享特征层和语音类别层的结构，如图 9.10 所示。语音模态块的学习是学习语音特征向量及其和原始语音数据的关系，视频模态块的学习是学习视频特征向量及其和原始视频数据的关系。两个块的学习方法一样，这里以语音模态块的学习为例，讲述其学习过程。给定 N 个输入数据结点，仍然从一个特征结点开始，然后应用结构 EM 学习算法，学习块结构，每次迭代不断加入新的特征结点，然后计算相应的块结构，这个过程直到获得的块结构评分不再提高为止。

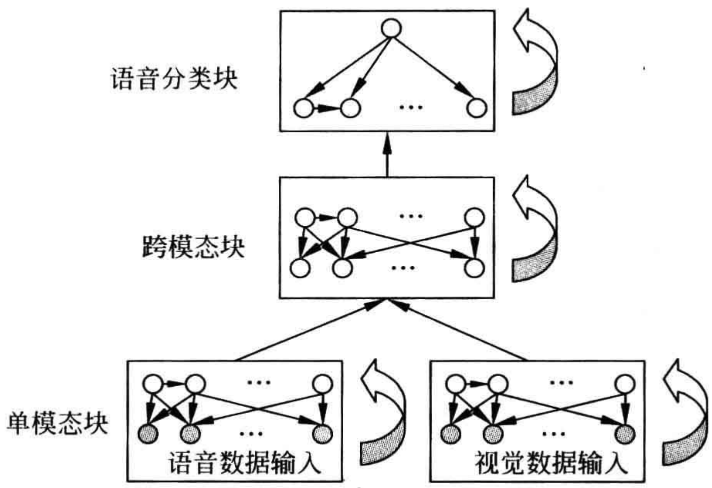


图 9.10 视听语音识别 DDBN 模型的空间网络学习

两个单模态块学习完成后，抽样语音特征变量和视频特征变量的数据，将其作为跨模态块中单模态特征层的输入数据。接着使用块学习方法，学习共享特征变量个数及其同语音特征和视频特征之间的链接关系。

在最上部的一个块，直接构建一个贝叶斯分类器，即语音类结点与各个共享特征结点完全链接。

2. 转移网络学习

空间网络学习完成后，假设 DDBN 模型在每一个时间片具有相同的空间结构，这样按照马尔可夫第一定律构建了 DDBN 的转移网络结构。由于一个 DDBN 可以看成是一个大的贝叶斯网络，因此使用贝叶斯网络参数 EM 学习算法来学习 DDBN 的参数。

3. 基于 DDBN 的视听语音识别模型学习算法

首先假设 $D=\{D_1,\cdots,D_M\}$ 代表 M 个视听语音时间序列数据，其中 D_m 是由 T 个 $D_{m,t}=\{a_{1,m}^t,\cdots,a_{P,m}^t,v_{1,m}^t,\cdots,v_{Q,m}^t,s_m^t\}$ 构成，其中 $1\leqslant t\leqslant T$，$a_{1,m}^t,\cdots,a_{P,m}^t$ 表示一个时间片的 P 个语音输入数据，$v_{1,m}^t,\cdots,v_{Q,m}^t$ 表示一个时间片的 Q 个视频输入数据，s_m^t 为该时间片的语音所属的语音类别。为了进行无监督学习，未标注语音数据和视频数据 $U_{_A(V)}=\{U_{_A(V)1},\cdots,$

$U_{_A(V)L}\}$需要被构建，其中$U_{_A(V)l}=\{a\ (v)_{1,l},\cdots,a\ (v)_{P(Q),l}\}\ (l=1,\cdots,L),L=\sum_{m=1}^{M}T_m$。$H^1_{_A(V)}=\{h^1_{_A(V)1},\cdots,h^1_{_A(V)k},\cdots\}$代表被学习到的单模态特征属性变量集，$H^2=\{h_1^2,\cdots,h_k^2,\cdots\}$代表共享特征变量集。

为了学习跨模态块的结构，$H^1_{_A(V)}$和H^2中的变量需要被推理计算，以便抽样出新的数据，给定学习到的单模态块结构，$H^1_{_A(V)}$中的变量可以按式(9.14)计算，即

$$h^{1,t^*}_{_A(V)_k}=\underset{h^{1,t}_{_A(V)k}}{\mathrm{argmax}}P(h^{1,t}_{_A(V)k}\mid \mathrm{Pa}(h^{1,t}_{_A(V)k}))\tag{9.14}$$

H^2中的变量可以按式(9.15)计算，即

$$h_k^{2,t*}=\underset{h_k^{2,t}}{\mathrm{argmax}}P(h_k^{2,t}\mid \mathrm{Pa}(h_k^{2,t}))\tag{9.15}$$

为了进行语音识别，使用标准的DBN推理算法，在给定观测数据$a_1^{1:t},\cdots,a_P^{1:t},v_1^{1:t},\cdots,v_Q^{1:t}$下，用式(9.16)获得$p(s^t\mid a_1^{1:t},\cdots,a_P^{1:t},v_1^{1:t},\cdots,v_Q^{1:t})$最大的$s^t$的值，即

$$s^{t*}=\underset{s^t}{\mathrm{argmax}}p(s^t\mid a_1^{1:t},\cdots,a_P^{1:t},v_1^{1:t},\cdots,v_Q^{1:t})\tag{9.16}$$

下面总结了给定输入数据D，基于DDBN的视听语音识别模型学习算法。

Algorithm9.2　The learning algorithm of DDBN for audiovisual speech recognition

1. Learning the audio block
   For i=1, … until no improvement in the score of the estimated structure between audio input layer and acoustic features layer
   1.1 Add one node into acoustic features layer
   1.2 Apply SEM to learn the structure using $U_{_A}$
2. Learning the visual block
   For i=1, … until no improvement in the score of the estimated structure between visual input layer and visual features layer
   2.1 Add one node into visual features layer
   2.2 Apply SEM to learn the structure using $U_{_V}$
3. Inferring the values of $H^1_{_A}$ and $H^1_{_V}$, according to the given data $U_{_A}$, $U_{_V}$ and the learned structures in step 1, 2
4. Learn the cross-modality block
   For i = 1, … until no improvement in the score of the estimated structure between the shared features layer and the single modality features layer
   4.1 Add one node into the shared features layer
   4.2 Apply SEM to learn the structure using $H^1=\{H^1_{_A}, H^1_{_V}\}$
5. Add the links between the label node and the shared feature nodes; (This step is not necessary for isolated-word recognition)
6. Construct the transition network topology of DDBN
7. Apply parameter EM to learn the parameters of the spatial network and transition network of DDBN using D
8. Output the DDBN models

9.3.4　实验与结果分析

这里使用了同9.2.4相同的藏语视听语音数据集，从每个数据中提取语音信号和视频

信号作为实验数据。语音信号被降采样到 8kHz，语音帧长 20ms，帧移 10ms，窗函数采用 Hamming 窗。因此对语音模态的输入层数据结点选为 160 个，其是一帧的原始语音信号幅度归一化处理后的数据。

对于视频数据，通过对每一帧图像进行处理，首先提取了说话者嘴部区域的 8 个点的位置，如图 9.11 所示，然后分别计算左右嘴角的距离和上下嘴唇对应点的距离，以代表说话者嘴的张度，如图 9.12 所示。由于每一帧图像中人脸的大小不同(图 9.13)，无法比较发音时嘴的动态变化，因此用相邻两帧的嘴的后差张度，即嘴部 4 个距离的前后两帧的差值作为当前帧的视频输入数据。并且由于视频信号的采样频率是 20Hz，为保证音视频输入数据的同步，对视频特征进行插值计算。

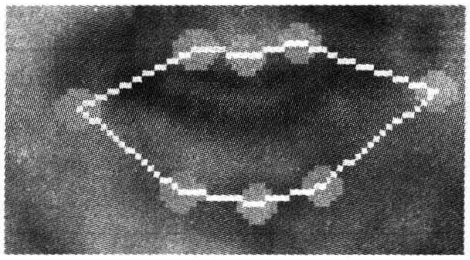


图 9.11 说话者嘴部区域的 8 个点

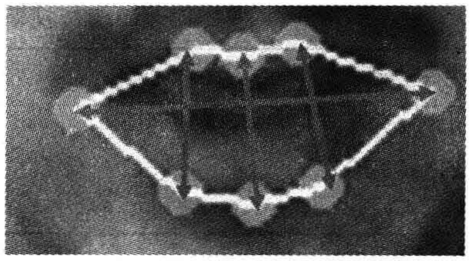


图 9.12 说话者嘴部张度的 4 个距离



图 9.13 3 个主题视听语音集中视频帧图像样本

同时又构造了无标注的数据集，用来无监督学习语音特征向量和共享特征向量，并且保证测试数据不包含在这个无标注数据集中。

根据先前的一些研究结果，即两个因子就可代表元音发音的稳定状态，假设隐连续状态变量为二维向量。

实验比较了基于 DDBN 的视听语音识别模型同两流的 DBN(图 9.14)和 Coupled HMM(图 9.15)的语音识别表现，其中两流的 DBN 和 Coupled HMM 的语音观测数据都取 24 维的 MFCC 特征参数，视频观测数据取嘴部的 4 个张度差。

表 9.2 示出了实验的比较结果，可以看到提出的 DDBN 用于视听语音识别，其识别结果远好于其他两个常用模型。这个模型可以很好地学习到语音特征向量和融合特征向量，

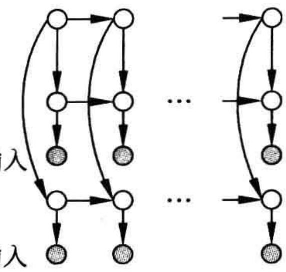


图 9.14 基于两流 DBN 的视听语音识别模型

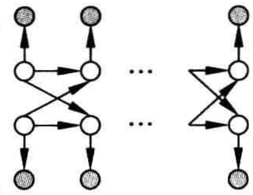


图 9.15 基于 Coupled HMM 的视听语音识别模型

更好地表示出了各模态信息的相关性和各属性之间的时空依赖关系。

表 9.2 视听语音识别模型实验结果

方法模型	Digits	Letters	Words
DDBN	90.1%	98.2%	99.4%
two-stream DBN	64.4%	72.3%	96.2%
Coupled HMM	62.3%	70.9%	95.4%

同时，也比较了视听语音识别模型同单模态语音识别模型的识别效果，结果如表 9.3 所示。可以看到，视听多模态融合识别结果要好于单模态识别结果。

表 9.3 双模态与单模态识别结果

方法模型	Digits	Letters	Words
Two-modal DDBN	90.1%	98.2%	99.4%
Audio-only DDBN	51.6%	66.7%	92.1%
Visual-only DDBN	53.4%	68.3%	88.6%

由于藏语是一个比较复杂的语言，其语言学至今还未研究成熟，因此对于藏语音素的研究成果还不能应用到藏语语音识别任务中。本节介绍的深度动态贝叶斯网络，其在无任何藏语先验知识的情况下，可以从语音数据中无监督地学习藏语语音特征，如果数据量足够大，其学习的特征变量接近于实际藏语语音音素个数。并且深度动态贝叶斯网络放松了帧间特征向量的独立假设，其更好地学习到了特征变量随时间变化的模式，通过对藏语 3 个主题的视听语音识别任务的实验可以看到，深度动态贝叶斯网络对视听融合语音识别的效果是非常好的。

附录A
概率图模型常用开发工具

多年来，研究者开发了多种软件和编程工具以支持概率图模型的学习和推理。下面列出了一些最常见的概率图模型软件开发工具。

1. BNT

贝叶斯网络工具箱是一个开源 Matlab 工具箱，针对有向 BNs（如高斯混合、线性/Logistic 回归、分层混合 experts、隐马尔可夫模型、线性动态系统及 DBNs）。BNT 包含联合树精确推理算法，它也支持诸如 Loopy BP 近似算法。对于 HMMs 模型，BNT 支持前向—后向推理和维特比推论。它也支持对 HMMs、BNs 和 DBNs 的最大似然估计（MLE）参数学习。BNT 也支持有限结点数目的 DBN 结构学习。其网站地址为 www. cs. ubc. ca/～murphyk /software/BNT/ bnt. html。PNL 是 BNT 的一个 C++ 版本，实现了部分 BNT 工具包的部分功能，支持在 MRF、BN、DBN 模型的推理。其网站地址 http://www. sourceforge. net/projects/openpnl。

2. MSBNx

MSBNx 是一个易于使用的图形用户界面软件。它支持离散 BN 模型的设计和精确推理。对于初学者，这个软件极其容易上手。它还支持 HMM 模型的概率推理。其网站地址为 http://research. microsoft. com/msbn/。

3. BUGS

BUGS(Bayesian Inference Using Gibbs Sampling)是一个用来通过 Gibbs Sampling 对 BNs 进行贝叶斯推理的软件。它对建模条件提供了大量的预编译分布。这个软件也很容易使用，因为它是基于脚本的界面。此外，BUGS 脚本可以被其他软件工具如 Matlab 直接调用。然而，对于抽样的 MCMC 收敛计算往往是非常缓慢的。其网站地址为 http://www. mrc-bsu. cam. ac. uk/bugs/。

4. VIBES

VIBES(Variational Inference Bayesian Networks)是一个 BN 推理的 Java 开源软件包。不像 BUGS 那样，它使用 Variational mean field 加快推理时间，但是却牺牲了模型的精确度。其网站地址为 http://vibes. sourceforge. net/。

5. Infer. NET

Infer. NET 是由 MicroSoft Research 开发的一个面向对象的概率图模型软件，用户可以指定(使用 Csoft 的)BNs 的一般模型图及具有离散/连续变量的因子图模型。它提供了丰富的学习和推理算法，包括广义信任传播和期望的传播。其网站地址为 http://research. microsoft. com/mlp/ml/infer/infer. htm。

6. CRF++

CRF++是一个用 C++编写的 CRF 工具包。它也为 Matlab 提供了 MAP 推理接口和边远概率计算的接口。CRF 模型仅限于 CRF 模型对。它对结点近邻限制在标准的 4-近邻系统。它支持使用 LBFGS 方法进行参数学习。网站地址为 http://crfpp. sourceforge. net/。

除了上述比较流行的工具包外，也有一些个人开发的软件包，如 CInference、SLS4MPE 和 libDAI。

附录B
贝叶斯网工具箱BNT的研究与学习

基于 Matlab 的贝叶斯网络工具箱 BNT 是 Kevin p. murphy 基于 Matlab 语言开发的关于贝叶斯网络学习的开源软件包,提供了许多贝叶斯网络学习的底层基础函数库,支持多种类型的结点(概率分布)、精确推理和近似推理、参数学习及结构学习、静态模型和动态模型。

贝叶斯网络表示:BNT 中使用矩阵方式表示贝叶斯网络,即若结点 i 到 j 有一条弧,则对应矩阵中 (i,j) 值为 1,否则为 0。

结构学习算法函数:BNT 中提供了较为丰富的结构学习函数,包括:

(1) 学习树扩展贝叶斯网络结构的 TAN 算法 learn_struct_tan()。

(2) 数据完整条件下学习一般贝叶斯网络结构的 K2 算法 learn_struct_k2()、贪婪搜索 GS(greedy search)算法 learn_struct_gs()和爬山 HC(hill climbing)算法 learn_struct_hc()等。

(3) 缺失数据条件下学习一般贝叶斯网络结构的最大期望 EM(Expectation Maximization)算法 learn_struct_EM()和马尔可夫链蒙特卡罗 MCMC(Markov Chain Monte Carlo)learn_struct_mcmc()算法等。

参数学习算法函数:BNT 中也提供了丰富的参数学习函数,都有:

(1) 完整数据时,学习参数的方法主要有两种:最大似然估计 learn_params()和贝叶斯方法 bayes_update_params()。

(2) 数据缺失时,如果已知网络拓扑结构,用 EM 算法来计算参数,倘若未知网络拓扑结构,使用结构最大期望 SEM(structure EM)算法 learn_struct_SEM()。

推理机制及推理引擎:为了提高运算速度,使各种推理算法能够有效应用,BNT 工具箱采用了引擎机制,不同的引擎根据不同的算法来完成模型转换、细化和求解。这个推理过程如下:

BNT 中提供了多种推理引擎,都有:

(1) 联合树推理引擎 jtree_inf_engine()。

(2) 全局联合树推理引擎 global_joint_inf_engine()。

(3) 信念传播推理引擎 belprop_inf_engine()。

(4) 变量消元推理引擎 var_elim_inf_engine()。

使用 BNT 做的相关实验及结果如下。

1. 手动建立贝叶斯网络模型及简单推理例子

例 B-1 “草地潮湿原因模型”建立及推理(全是离散变量)。

```
%建立贝叶斯网络结构并制定条件概率表
N = 4;  %四个结点 分别是 cloudy, sprinkler, rain, wetgrass
dag = zeros(N,N);
C = 1; S = 2; R = 3; W = 4;
dag(C,[R S]) = 1;              %结点之间的连接关系
dag(R,W) = 1;
dag(S,W) = 1;
discrete_nodes = 1:N;          %离散结点
node_sizes = 2 * ones(1,N);    %结点状态数
bnet= mk_bnet(dag,node_sizes,'names',{'cloudy','sprinkler','rain','wetgrass'},'discrete',discrete
_nodes);
bnet.CPD{C} = tabular_CPD(bnet,C,[0.5 0.5]);
bnet.CPD{R} = tabular_CPD(bnet,R,[0.8 0.2 0.2 0.8]);
bnet.CPD{S} = tabular_CPD(bnet,S,[0.5 0.9 0.5 0.1]);
bnet.CPD{W} = tabular_CPD(bnet,W,[1 0.1 0.1 0.01 0 0.9 0.9 0.99]);
%画出建立好的贝叶斯网络
figure
draw_graph(dag)
%选择 jtree_inf_engine 推理引擎.
engine = jtree_inf_engine(bnet);
%输入证据
evidence = cell(1,N);
evidence{R} = 2;
[engine, loglike]  = enter_evidence(engine, evidence);
%计算单个结点后验概率,即进行推理
marg1 = marginal_nodes(engine,S);
marg1.T
%计算对结点联合后验概率
marg2 = marginal_nodes(engine,[S R W]);
marg2.T
%给出“软证据”,即结点的可能分布概率情况下的推理
evidence{R} = [];
soft_evidence{R}=[0.6 0.4];
[engine, loglike]  = enter_evidence(engine,evidence,'soft',soft_evidence);
marg3 = marginal_nodes(engine,S);
marg3.T
```

运行结果:

(1) 贝叶斯网络如图 B-1 所示。

(2) 单个节点后验概率:

```
ans =
   0.8200
   0.1800
```

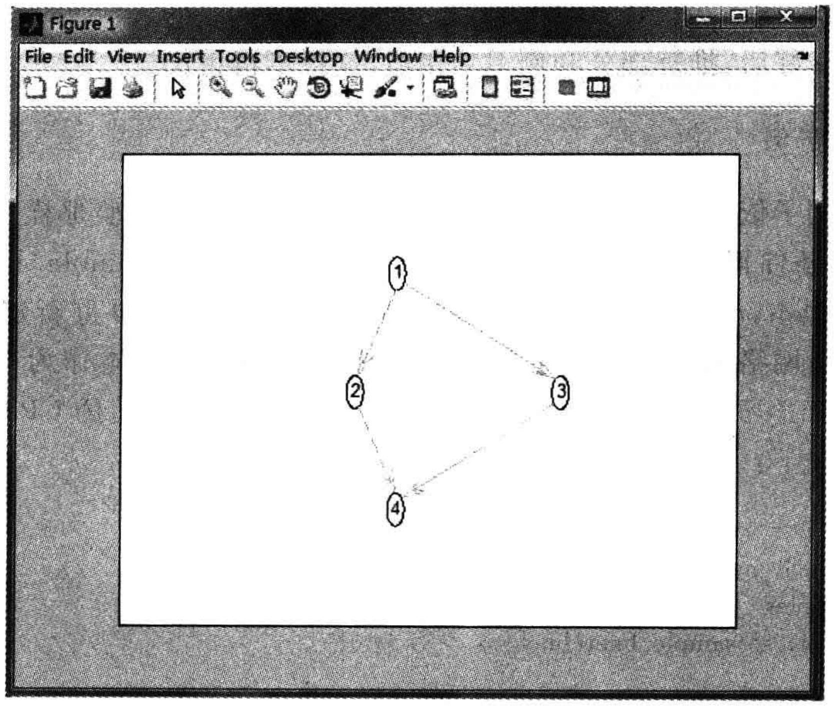


图 B-1　草地潮湿原因贝叶斯网络模型

（3）多个节点后验概率：

```
ans(:,:,1) =
   0.0820
   0.0018
ans(:,:,2) =
   0.7380
   0.1782
```

（4）soft_evidence 情况下的后验概率：

```
ans =
   0.6760
   0.3240
```

例 B-2　焚化炉厂废物排放模型建立及推理(包含离散变量和连续变量)。

这个实验与第一个实验不同的地方就是它所建立的贝叶斯网络中的结点变量包含连续变量，在建立条件概率表时会有所不同，离散变量使用 CPD 构造器 tabular_CPD，连续变量使用 gaussian_CPD。这里只给出这一部分的代码：

```
bnet.CPD{B} = tabular_CPD(bnet,B,'CPT',[0.85 0.15]);
bnet.CPD{F} = tabular_CPD(bnet,F,'CPT',[0.95 0.05]);
bnet.CPD{W} = tabular_CPD(bnet,W,'CPT',[2/7 5/7]);
bnet.CPD{E} = gaussian_CPD(bnet,E,'mean',[-3.9 -0.4 -3.2 -0.5],'cov',[0.00002 0.0001
0.00002 0.0001]);
bnet.CPD{D} = gaussian_CPD(bnet,D,'mean',[6.5 6.0 7.5 7.0],'cov',[0.03 0.04 0.1 0.1],'
weights',[1 1 1 1]);
bnet.CPD{C} = gaussian_CPD(bnet,C,'mean',[-2 -1],'cov',[0.1 0.3]);
bnet.CPD{L} = gaussian_CPD(bnet,L,'mean',3,'cov',0.25,'weights',-0.5);
bnet.CPD{Min} = gaussian_CPD(bnet, Min,'mean',[0.5 -0.5],'cov',[0.01 0.005]);
```

```
bnet.CPD{Mou}=gaussian_CPD(bnet,Mout,'mean',0,'cov',0.002,'weights',[1 1]);
```

贝叶斯网建立好后，推理部分和例 B-1 相似，这里不再讨论。

2. 参数学习举例

这里使用的例子依然是“草地潮湿原因模型”。首先如上面实验那样建立好贝叶斯网络 bnet，并手动构造条件概率表 CPT。然后使用 BNT 里的函数 sample_bnet(bnet)来产生 nsamples 个数据样本，nsamples 分别取值 20、200、2000。然后，再重新建立一个不知道条件概率表的贝叶斯网络 bnet2(结构和 bnet 相同)，并把得到的样本作为训练集代入 learn_params()函数进行学习，把学习到的条件概率表 CPT2 与手动构造的 CPT 进行了比较。参数学习部分代码如下：

```
nsamples = 20;
samples = cell(N, nsamples);
for i = 1:nsamples
    samples(:,i) = sample_bnet(bnet);
end
data = cell2num(samples);
bnet2 = mk_bnet(dag,node_sizes,'discrete',discrete_nodes);
seed = 0;
rand('state',seed);
bnet2.CPD{C} = tabular_CPD(bnet2,C);
bnet2.CPD{S} = tabular_CPD(bnet2,S);
bnet2.CPD{R} = tabular_CPD(bnet2,R);
bnet2.CPD{W} = tabular_CPD(bnet2,W);
bnet3 = learn_params(bnet2,data);
CTP 1=cell(1,N);
for i=1:N
  s=struct(bnet3.CPD{i});
  CPT 1{i}=S.CPT;
end
```

运行结果：

手动给出的 CPT

```
  [0.5 0.5]
  [0.8 0.2 0.2 0.8]
  [0.5 0.9 0.5 0.1]
  [1 0.1 0.1 0.01 0 0.9 0.9 0.99]
nsamples=20
  [0.4 0.6]
  [0.6250 0.1666 0.3750 0.8333]
  [0.5 1 0.5 0]
  [1 0.09 0 0 0 0.9091 1 1]
nsamples=200
  [0.51 0.49]
  [0.7451 0.1735 0.2549 0.8265]
  [0.4608 0.908 0.5392 0.0918]
  [1 0.1047 0.0930 0 0 0.8953 0.9070 1]
```

```
nsamples=2000
  [0.5085 0.4915]
  [0.7906 0.2094 0.2014 0.7986]
  [0.4749 0.8932 0.5251 0.1068]
  [1 0.1023 0.0796 0.0053 0 0.8977 0.9204 0.9947]
```

可以看出，随着训练样本数的增加，学习到的条件概率表越来越逼近于手动给出的条件概率表。

3. 结构学习举例

基于上面“草地潮湿原因模型”例子，首先手动建立好贝叶斯网络 bnet，然后使用函数 sample_bnet(bnet)产生训练样本，制定结点顺序和最大父结点数，代入 K2 算法，进行结构学习，并比较了不同训练样本数(nsamples)的情况下，学习到的结构和实际结构的差异。

添加的程序代码如下：

```
order = [C S R W];
max_fan_in =2;
dag2=learn_struct_K2(data, node_sizes, order, 'max_fan_in', max_fan_in);
figure
draw_graph(dag2)
```

实验结果如下：

可以看出当样本数达到一定的值以后，K2 算法可以很好地建立出需要的贝叶斯网络结构。

(1) 手动给出的结构，如图 B-2 所示。

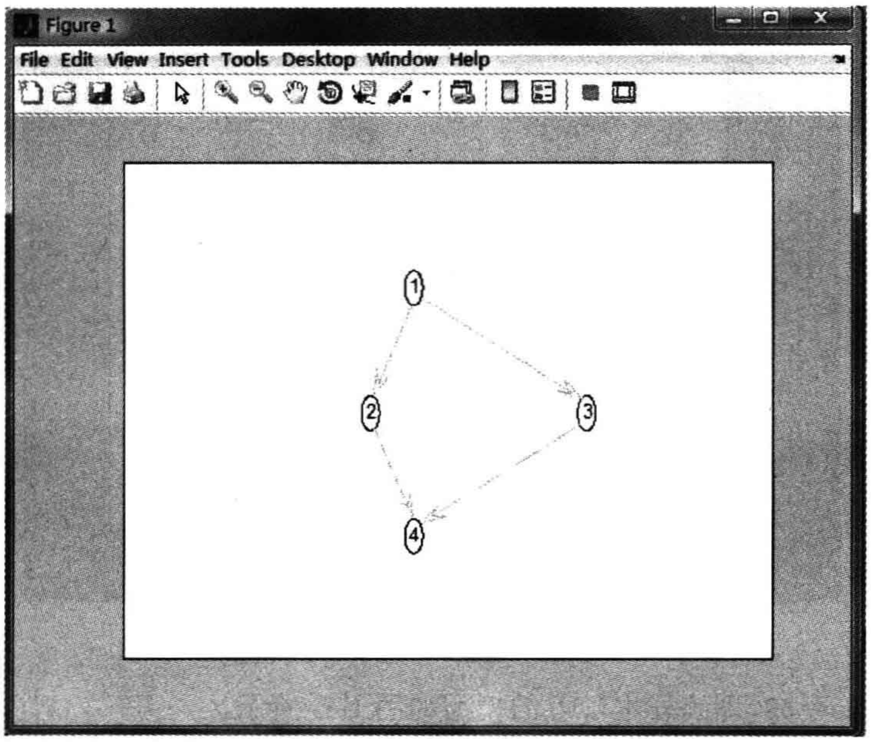


图 B-2　手动给出的贝叶斯网络模型

（2）nsamples=10，训练样本学习的模型如图 B-3 所示。

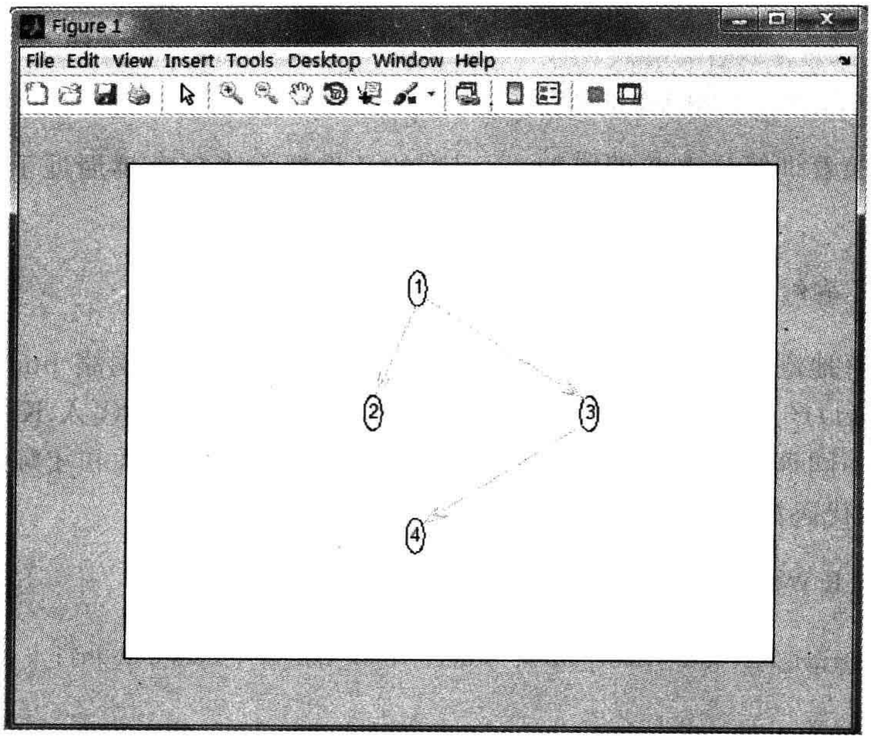


图 B-3 10 个训练样本学习的贝叶斯网络模型

（3）nsamples=20，训练样本学习的模型如图 B-4 所示。

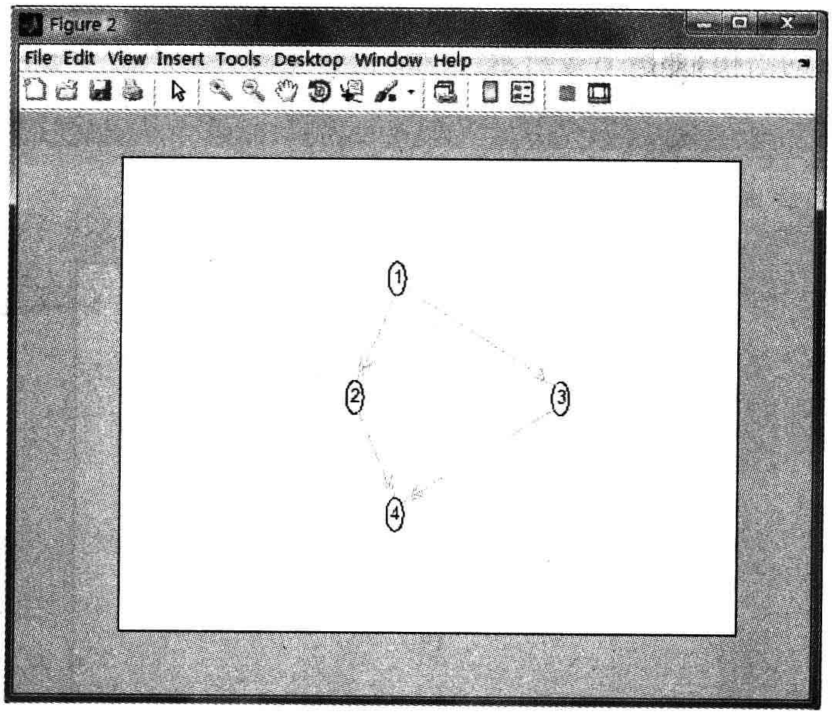


图 B-4 20 个训练样本学习的贝叶斯网络模型

（4）nsamples=30，训练样本学习的模型如图 B-5 所示。

4. *应用实例*

信用风险等级评估模型学习和预测。

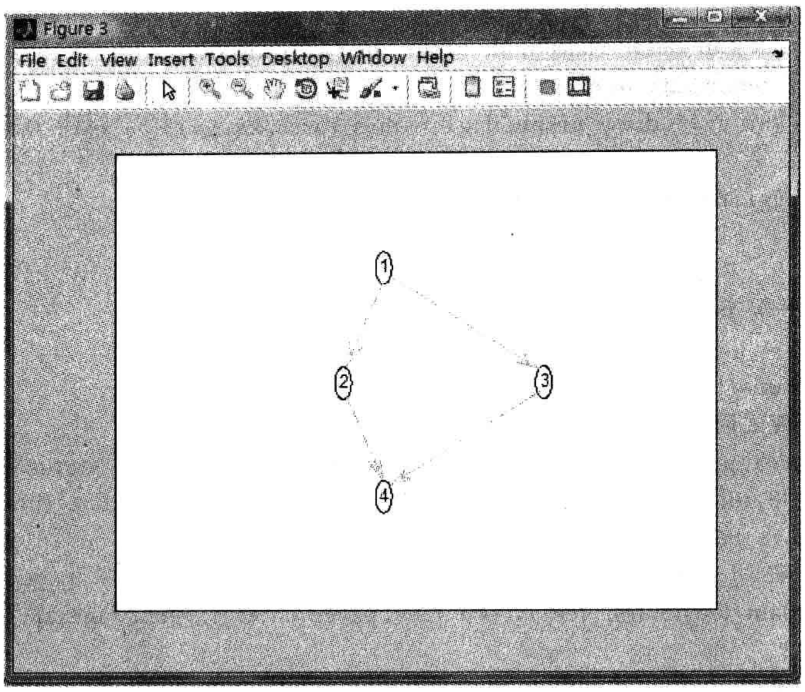


图 B-5　30 个训练样本学习的贝叶斯网络模型

本应用程序对客户的 15 个特征属性建立信用风险等级分类模型，模型结点共 16 个，包含一个等级分类结点 LIMIT_OWE_L，结点全部为离散变量。分类模型采用 TAN 模型，通过对输入的训练集进行结构和参数学习，建立分类模型，最后对其进行分类精度测试。下面是其完整代码。

```
%初始化变量和读数据
N=16;
PAY_CODE=1;D_CODE=2;PREPAY_FEE_L=3;FAVOUR_FEE_L=4;SHOULD_PAY_L=5;
PAYED_STATUS=6;MONTH_FEE_L=7;LOCAL_NET_L=8;ROAM_FEE_L=9;
TOLL_FEE_L=10;OCC_FEE_L=11;PAYED_LATER_L=12;SEX=13;AGE_L=14;LIMIT_
OWE_L=16;ONLINE_TIME_L=15;
ns = [2 2 5 3 5 2 3 4 4 5 2 5 3 5 4 4]; %每个结点的取值范围
seed = 0;
rand('state', seed);
randn('state', seed);
%读训练数据
[pay_code,d_code,prepay_fee_l,favour_fee_l,should_pay_l,payed_status,month_fee_l,local_net_l,
roam_fee_l,toll_fee_l,occ_fee,payed_later_l,sex,age_l,online_time_l,limit_owe_l]=textread('limit_
owe_alltrain.txt','%d%d%d%d%d%d%d%d%d%d%d%d%d%d%d%d');
data=[pay_code,d_code,prepay_fee_l,favour_fee_l,should_pay_l,payed_status,month_fee_l,local_
net_l,roam_fee_l,toll_fee_l,occ_fee,payed_later_l,sex,age_l,online_time_l,limit_owe_l]';
[n1 n1_case]=size(data);
order=[PAY_CODE D_CODE PREPAY_FEE_L FAVOUR_FEE_L SHOULD_PAY_L PAYED_
STATUS MONTH_FEE_L LOCAL_NET_L ROAM_FEE_L TOLL_FEE_L OCC_FEE_L PAYED_
LATER_L SEX AGE_L ONLINE_TIME_L LIMIT_OWE_L];
max_fan_in = N;
class_node=16;
root=15;
scoring_fn='bic';
```

```
%结构学习
dag1 = learn_struct_tan(data, class_node, root, ns, scoring_fn);
figure; [xx yy]= make_layout(dag1);
subplot(1,2,1), [xx yy]=draw_graph(dag1,names,carre,xx,yy); %,carre);
title('dag1');
bnet=mk_bnet(dag1,ns,'discrete',order);
% 参数学习
For i=1:16
    bnet.CPD{i} = tabular_CPD(bnet,i);
end
bnet = learn_params(bnet,data);
%读测试数据并测试模型分类精度
[pay_code3,d_code3,prepay_fee_l3,favour_fee_l3,should_pay_l3,payed_status3,month_fee_l3,local_net_l3,roam_fee_l3,toll_fee_l3,occ_fee3,payed_later_l3,sex3,age_l3,online_time_l3,limit_owe_l3]=textread('limit_owe_test.txt','%d%d%d%d%d%d%d%d%d%d%d%d%d%d%d%d');
data_t=[pay_code3,d_code3,prepay_fee_l3,favour_fee_l3,should_pay_l3,payed_status3,month_fee_l3,local_net_l3,roam_fee_l3,toll_fee_l3,occ_fee3,payed_later_l3,sex3,age_l3,online_time_l3,limit_owe_l3]';
[tt ttcase]=size(data_t);
pv_t=zeros(1,ttcase);
p_t=zeros(1,ttcase);
engine_t = jtree_inf_engine(bnet);
pv=zeros(1,ttcase);
p=zeros(1,ttcase);
pc=zeros(1,4);
i_class=0;
count=0;
L11=0;L12=0;L13=0;L14=0;nc1=0;
L21=0;L22=0;L23=0;L24=0;nc2=0;
L31=0;L32=0;L33=0;L34=0;nc3=0;
L41=0;L42=0;L43=0;L44=0;nc4=0;
for i=1:ttcase
evidence = cell(1,N);
    for j=1:N-1
        evidence{j} = data_t(j,i);
    end
[engine_t, ll_t]  = enter_evidence(engine_t, evidence);
m = marginal_nodes(engine_t, class_node);
    for k=1:4
        pc(k)= m.T(k);
    end
max_pc=pc(1);
max_i=1;
r_m=zeros(1,4);
for f=1:3
    if max_pc>pc(f+1),
        r_m(f)=0;
    else
        if max_pc==pc(f+1),
            r_m(f)=1;
        else
```

```
                max_pc=pc(f+1);
                max_i=f+1;
                r_m(f+1)=0;
            end
        end
    end
    if r_m(max_i)==1,
       i_class=i_class+1;
    else
       p(i)=max_pc;
       pv(i)=max_i;
    end
     v=data_t(N,i);
    if v==1&pv(i)==1,count=count+1 ;L11=L11+1;end
    if v==1&pv(i)==2,L12=L12+1;end
    if v==1&pv(i)==3,L13=L13+1;end
    if v==1&pv(i)==4,L14=L14+1;end
    if v==2&pv(i)==1,L21=L21+1;end
    if v==2&pv(i)==2,count=count+1 ;L22=L22+1;end
    if v==2&pv(i)==3,L23=L23+1;end
    if v==2&pv(i)==4,L24=L24+1;end
    if v==3&pv(i)==1,L31=L31+1;end
    if v==3&pv(i)==2,L32=L32+1;end
    if v==3&pv(i)==3,count=count+1 ;L33=L33+1;end
    if v==3&pv(i)==4,L34=L34+1;end
    if v==4&pv(i)==1,L41=L41+1;end
    if v==4&pv(i)==2,L42=L42+1;end
    if v==4&pv(i)==3,L43=L43+1;end
    if v==4&pv(i)==4,count=count+1 ;L44=L44+1;end
    if r_m(max_i)==1&v==1,nc1=nc1+1;end
    if r_m(max_i)==1&v==2,nc2=nc2+1;end
    if r_m(max_i)==1&v==3,nc3=nc3+1;end
    if r_m(max_i)==1&v==4,nc4=nc4+1;end
    end
    per_matrix=[];
    per_matrix(1)=count/ttcase; %分类精度
```

参 考 文 献

[1] 史忠植. 知识发现[M]. 北京：清华大学出版社，2002.

[2] 卢增祥，李衍达. 交互支持向量机学习算法及其应用[J]. 清华大学学报（自然科学版），1999，39(7)：93-97.

[3] Almgren M, Jonsson E. Using Active Learning in Intrusion Detection [A]. In: Proceedings of the 17th IEEE on Computer Security Foundations Workshop[C], Januarary 2004, 88-98.

[4] He J R, Li M J, Zhang H J. Mean Version Space: a New Active Learning method for Content-Based Image Retrieval [A]. In: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval[C]. New York, USA, 2004, 15-22.

[5] 孙功星. 神经网络主动学习的进化算法[J]. 计算机科学，2002，29(10)：61-63.

[6] 周平. 遗传程序设计及其新进展与应用[J]. 浙江工程学院学报，2003，20(1)：21-25.

[7] 刘宏. 一种新的支持向量机主动学习策略及其在文本分类中的应用[J]. 计算机科学，2003 ，30(6)：110-114.

[8] 张健沛，徐华. 支持向量机(SVM)主动学习方法研究与应用[J]. 计算机应用，2004，24(1)：1-3.

[9] Mitra P, Murthy C A, Pal S K. A Probabilistic Active Support Vector Learning Algorithm[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2004, 26(3): 413-418.

[10] Melville P, Mooney R J. Diverse Ensembles for Active Learning[A]. In: Proceedings of the 21st International Conference on Machine Learning[C]. Banff, Canada, July 2004, 584-591.

[11] Yi X, Kou Z B, Zheng C S. Classifier Combination Based on Active Learning [A]. In: Proceedings of the 17th International Conference on Pattern Recognition[C]. Los Alamitos, USA, August 2004, 184-187.

[12] 徐杰，施鹏飞. 图像检索中基于标记与未标记样本的主动学习算法[J]. 上海交通大学学报，2004，38(12)：2068-2072.

[13] 丁欣. 国外信用风险评估方法的发展现状[J]. 湖南大学学报(社会科学版)，2002，5：140-142.

[14] 宫秀军，孙建平，史忠植. 主动贝叶斯网络分类器[J]. 计算机研究与发展，2002，39(5)：574-579.

[15] 覃刚力. 基于主动学习的文档分类[J]. 计算机科学，2003，30(10)：45-48.

[16] 徐继生，王道恒. 移动通信用户信用度的预测评估[J]. 电信科学，2002，(8)：17-20.

[17] 张玉洁，孟祥武. 基于遗传算法的电信客户初始信用度分配算法[J]. 北京邮电大学学报，2002，25(2)：74-77.

[18] 中国移动通信集团公司. 中国移动经营分析系统业务规范(V2.0)[S]. 北京：中国移动通信集团公司，2002.

[19] 孟阳. 信用度量制模型与商业银行信用风险量化度量管理[J]. 现代财经，2003，23(3)：32-34.

[20] 杨奕. 电信企业如何建立信用度管理制度[J]. 电信科学，2002，(2)：51-52.

[21] 胡玉胜. 有环·动态 Bayes 网络研究及其应用. 博士论文. 北京科技大学，2001.

[22] 田凤占，陆玉昌. 用动态 Bayesian 网络建立宏观经济系统模型[J]. 清华大学学报(自然科学版)，2004，44(9)：1256-1259.

[23] Murphy K P. Dynamic Bayesian Networks: Representation, Inference and Learning. PHD Thesis. University of California, Berkeley, 2002.

[24] Lewis D D, Gale W A. A Sequential Algorithm for Training Text Classifiers [A]. In: Proceedings of 17th ACM International Conference on Research and Development in Information Retrieval [C]. Dublin, Ireland, July 1994, 3-12.

[25] Argamon-Engleson S, Dagan I. Committee-based Sample Selection for Probabilistic Classifiers [J]. Journal of Artificial Intelligence Research, 1999, 11: 335-460.

[26] Freund Y, Seung H S, Samir E, Tishby N. Selective Sampling Using the Query By Committee Algorithm[J]. Machine Learning, 1997, 28(2-3):133-168.

[27] McCallum A K, Nigam K. Employing EM and Pool-based Active Learning for Text Classification [A]. Proceeding of the 15th International Conference on Machine Learning[C]. Madison, USA, July1998, 350-358.

[28] Cohn D A, Ghahramani Z, Jordan M I. Active Learning with Statistical Models [J]. Journal of Artificial Intelligence Research, 1996, 4:129-145.

[29] Seung H S, Opper M, Sompolinsky H. Query by Committee[A]. Proceedings of the 15th Annual ACM Workshop on Computational Learning Theory[C]. Pittsburgh, USA, July 1992, 287-294.

[30] Riccardi G. Active Learning: Theory and Applications to Automatic Speech Recognition [J]. IEEE Transaction on Speech and Audio Processing, 2005, 13(4):504-511.

[31] Abe N, Mamitsuka H. Query Learning Strategies Using Boosting and Bagging[A]. In: Proceedings of the Fifteenth International Conference Machine Learning. (ICML'98)[C]. Madison, USA, July 1998, 1-9.

[32] 邢永康. 信度网理论及其应用研究. 博士论文. 重庆大学, 2001.

[33] 宫秀军，刘少辉，史忠植.一种增量贝叶斯分类模型[J]. 计算机学报,2002,25(6):645-650.

[34] 李建中.用于分类规则挖掘的贝叶斯信念构造算法. http://www.bjx.com.cn/files/wx/jsjyxxjs/2002-5/1.htm, 2002.

[35] 周颜军,王双成,王辉.基于贝叶斯网络的分类器研究[J].东北师大学报自然科学版,2003,35(2):21-27.

[36] 林士敏,王双成,陆玉昌. Bayesian 方法的计算学习机制和问题求解[J]. 清华大学学报(自然科学版), 2000, 40(9):61-64.

[37] 冀俊忠，刘椿华，沙志强. 贝叶斯网模型的学习、推理和应用[J]. 计算机工程与应用,2003,(5):24-27.

[38] 张宏伟,田凤占,陆玉昌. 对一种贝叶斯网络学习算法的改进和试验性分析[J]. 计算机科学,2002,29(5):97-100.

[39] 王双成,林士敏,陆玉昌.不完整数据的贝叶斯网络学习问题[J]. 清华大学学报, 2000,40(9):65-68.

[40] 林士敏，田凤占，陆玉昌. 贝叶斯网络的建造及其在数据采掘中的应用[J]. 清华大学学报, 2000, 40(1):49-52.

[41] 林士敏，田凤占，陆玉昌. 贝叶斯学习、贝叶斯网络与数据采掘[J]. 计算机科学, 2000,27(10): 69-72.

[42] 林士敏，田凤占，陆玉昌. 用于数据采掘的贝叶斯分类器研究[J]. 计算机科学, 2000, 27(10): 73-76.

[43] 王双成，林士敏，陆玉昌.贝叶斯网络结构学习分析. 计算机科学, 2000,27(10): 77-79.

[44] 王双成，林士敏，陆玉昌. 用贝叶斯网络进行因果分析[J]. 计算机科学, 2000, 27(10): 80-82.

[45] Lam W, Bacchus F. Learning Bayesian Belief Networks: an Approach Based on the MDL Principle [J]. Computational Intelligence, 1994, 10(3):269-293.

[46] Cooper G F, Herskovtis E. A Bayesian Method for the Induction of Probabilistic Network from Data [J]. Machine Learning, 1992, (9): 309-347.

[47] Friedman N, Yakhini M. On the Sample Complexity of Learning Bayesian Networks [A]. In: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence[C]. Portland, USA, August 1996, 274-282.

[48] Heckman D, Shachter R. Decision-Theoretic Foundations for Causal Reasoning [J]. Journal of Artificial Intelligence Research, 1995, 7(3):405-430.

[49] Shachter R D. Probabilistic Inference in Influence Diagrams [J]. Operations Research, 1998, 36(4): 589-604.

[50] Herskouts E. Computer-based Probabilistic Network Construction [D]. Docloral dissertation. Medical information sciences. Stanford University. Stanford, CA, 1991.

[51] Suzuki J. Learning Bayesian Belief Networks Based on the MDL Principle: an Efficient Algorithm Using the Branch and Bound Technique [A]. In: Proceedings of the international conference on machine learning[C]. Bari, Italy, July 1996, 462-470.

[52] Preparata F, Shamos M. Computational Geometry: an Introduction [M]. New York: Springer-Verlag, 1988.

[53] Dean T, Kanazawa K. A Model for Reasoning about Persistence and Causation [J]. Computational Intelligent, 1989, 5(3): 142-150.

[54] Murphy K P. A Brief Introduction to Bayes' Rule. http://www.ai.mit.edu/~murphyk, 2000.

[55] Liu W Y, Song N. Fuzzy Functional Dependencies and Bayesian Networks [J]. J. Comput. Sci. & Technol, 2003, 18(1):56-66.

[56] Cheng J, Bell D A, Liu W. Learning Belief Network from Data: An Information Theory Based Approach [A]. In: Proceedings of ACM CIKM'97[C], 1997, 325-331.

[57] Murphy K. Active Learning of Causal Bayes Net Structure. http://www.ai.mit.edu/~murphy/publ.html, 2001, 3.

[58] Tong S, Koller D. Active Learning for Structure in Bayesian Networks [A]. In: Proceedings of Seventeenth International Joint Conferrence on Artificial Intelligence [C]. Seatlle, Washington, August 2001, 863-869.

[59] Geiger D, Heckerman D. A Characterization of Dirichlet Distribution with Applicable to Learning Bayesian networks [A]. In: Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence [C]. Montreal, Canada, 1995, 196-207.

[60] Dempster P, Laird N M, Rubin D B. Maximum Likelihood From Incomplete Data Via the EM Algorithm[J]. Royal Stat. Soc., 1997, 39(1):1-38.

[61] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers [J]. Machine Learning, 1997, 29(2-3):131-161.

[62] Eamonn E J, Pazzani M J. Learning Augmented Bayesian Classifiers: a Comparison of Distribution Based and Classification Based Approaches [A]. In: Proceedings of the 7th Int'l Workshop on Artificial Intelligence and Statistics[C]. 1999, 225-230.

[63] Madden G M. Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm [R]. Technical Report NO. NUIG-IT-011002. Department of information Technology, National University of Ireland, Galway, 2002.

[64] 田凤占，张宏伟，陆玉昌，田凤奇. 动态贝叶斯网络的学习和近似推断[A]. In: Proceedings of the 4th World Congress on Intelligent Control and Automation [C], Shanghai, China, June 2002, 2291-2296.

[65] 王飞，刘大有，卢奕南. 基于遗传算法的动态 Bayesian 网结构学习的研究[J]. 电子学报，2001，38(8): 698-702.

[66] 姚宏亮，王浩，胡学刚. 基于动态贝叶斯网络的多 Agent 交互的模型表示[J]. 计算机工程，2003，29(16): 57-59.

[67] 王天树，郑南宁，袁剑泽. 机器智能与模式识别研究中的统计学习方法[J]. 自动化学报，2002，28 Suppl: 103-116.

[68] Friedman N, Murphy K, Russell S. Learning the Structure of Dynamic Probabilistic Networks [A]. In:Proceedings of Uncertainty in Artificial Intelligence (UAI-98) [C], Madison, USA, July 1998,

139-147.

[69] Friedman N. Learning Belief Networks in the Presence of Missing Values and Hidden Variables [A]. In: Proceedings of the 14th International Conference on Machine Learning[C]. Nashville, USA, 1997, 125-133.

[70] 徐杰. 基于小样本学习的图像检索研究. 博士论文. 上海交通大学,2004.

[71] Duda R O, Hart P E, Stork D G. Pattern classification[M]. Second Edition, John Wiley & Sons, Inc. , 2001.

[72] Roy N, McCallum A. Toward Optimal Active Learning through Sampling Estimation of Error Reduction [A]. In: proceedings of the 18th Int'1 Conf on Machine Learning (ICMI-2001) [C]. Western Massachusetts, 2001, 441-448.

[73] Souvannavong F, Merialdo B, Huet B. Partition Sampling: an Active Learning Selection Strategy for Large Database Annotation [A]. In: IEE Proc. -Vis. Image Signal Process[C]. 2005,347-355.

[74] Gale W A, Church K W. A Program for Aligning Scentence in Bilingual Corpora [J]. Computational linguistics, 1993, 19(1):75-102.

[75] Mitchell T M. 机器学习[M]. 北京:机械工业出版社, 2003.

[76] Tong S, Koller D. Active Learning for Parameter Estimation in Bayesian Networks [A]. In: Proc. NIPS 13, 2001.

[77] 傅祖芸. 信息论基础[M]. 北京:电子工业出版社, 1986.

[78] 白鹏. 两个一元 t-分布之间的 Kullback-Leibler 距离[J]. 数学物理学报,2002,22A(1): 121-127.

[79] Kim S Y, Imoto S, Miyano S. Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data[J]. Computational Methods in Systems Biology,2003: 104～113.

[80] Gang A, Pavlonic V. Boosted Learning in Dynamic Bayesian Networks for Multionodal Speaker Detection [A]. In: Proceedings of the [C], September 2003, 91(9):1355-1369.

[81] 李庆中,苑春法,黄锦辉. 基于小规模标注语料的机器学习方法研究[J]. 计算机应用, 2004, 24 (2):56-58.

[82] Cohen I, Cozman F G, Sebe N. Semi-supervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Conputer interaction [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2004, 26(12):1553-1567.

[83] 宫秀军,史忠植. 基于 Bayes 潜在语义模型的半监督 Web 挖掘[J]. 软件学报, 2002,13(8): 1508-1514.

[84] Yan R, Hauptmann A. Multi-Class Active Learning for Video Semantic Feature Extraction[A]. In: IEEE International Conference on Multimedia and Expo(ICME)[C]. Taipei, 2004, 69-72.

[85] Tong S, Chang E. Support Vector Machine Active Learning for Image Retrieval[A]. In Proceedings of 9th ACM Multimedia Conference[C]. Canada, 2001,107-118.

[86] Kothari R, Jain V. Learning From labeled and Unlabeled Data Using a Minimal Number of Queries [J]. IEEE Transaction on Neural Networks, 2003, 14(6):1496-1505.

[87] 郭人人,梁世栋,方兆本. 消费者信用评估分析综述[J]. 系统工程,2001,19(6):9-15.

[88] Thomas L C. A Survey of Credit and Behavioural Scoring Forcasting Financial Risk of Lending to Consumers [J]. International Journal of Forecasting, 2000, 16: 149-172.

[89] 李健云,邱菀华. 核 Fisher 判别分析方法评估消费者信用风险[J]. 系统工程理论方法应用, 2004, 13(6): 548-552.

[90] Baesens B, Egmont-petersen M, Castelo R, Vantlienen J. Learning Bayesian network Classifiers for Credit Scoring Using Markov Chain Monte Carlo search[A]. In: Proceedings of 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3[C]. Quebec Canada, August 2002,49-52.

[91] Han J, Kamber M. Data Ming Concepts and Techniques [M]. 北京：机械工业出版社，2001.

[92] Smith A, Elkan C. A Bayesian Networks Framwork for Reject Inference[A]. In: KDD'04[C], Seattle, USA, August 2004, 286-295.

[93] 刘同明. 数据挖掘技术及其应用[M]. 北京：国防工业出版社，2001.

[94] 罗宁，穆志纯. 基于贝叶斯网的分类器及其在 CRM 中的应用[J]. 计算机应用，2004，24(3)：79-81.

[95] 冀俊忠，刘椿年，沙志强. 贝叶斯网络模型的学习、推理和应用[J]. 计算机工程与应用，2003，39(5)：24～27.

[96] 王飞. Bayesian 网学习算法的研究. 博士论文. 吉林大学，2001.

[97] 张少中，王秀坤，孙莹光. 贝叶斯网络及其在决策支持系统中的应用[J]. 计算机工程，2004，30(10)：1-3.

[98] 陈俊艳，徐继生，李湛. BP 神经网络技术在移动通信客户信用分析中的应用[J]. 武汉大学学报(理学版)，2005，51(1)：114-118.

[99] 田辽，张克平，覃征. 贝叶斯构造网络算法在电信 CRM 中的应用[J]. 现代电子技术，2005，201(10)：52-54.

[100] 中国联通客服、结算与信息系统部. 中国联通统一经营信息服务平台业务规范(讨论稿)[S]. 北京：中国联通客服、结算与信息系统部，2002.

[101] 王珊. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1999.

[102] 廖理，余英泽，吴瑜，聂能. 数据挖掘和数据仓库及其在电信业中的应用[J]. 重庆邮电学院学报，2000，12(4)：31-34.

[103] 罗宁，穆志纯. CRM 在电信行业中的应用与技术[J]. 计算机时代，2004，(1)：45～46.

[104] 赵瑞锋，孟祥武. 数据仓库技术及其在电信反欺诈中的应用[J]. 计算机应用研究，2004，(2)：77-79.

[105] 赵宏波. 电信企业客户关系管理[M]. 北京：人民邮电出版社，2003.

[106] 赵悦，周利民，付冬梅. 基于数据仓库和 OLAP 的移动通信决策支持系统[J]. 计算机系统应用，2002，(4)：10-13.

[107] 叶进，程泽凯，林士敏. 基于贝叶斯网络的电信客户流失预测模型[J]. 计算机工程与应用，2005，14：212-214.

[108] Zhao Y, Li B, Li X, Liu W H. Customer Churn Prediction Using Improved One-Class Support Vector Machine[A]. In: Proceedings of First International Conference on Advanced Data Mining and Applications[C]. Wuhan, China, July 2005, 300-306.

[109] 罗宁. OLAP 技术及贝叶斯网在电信 CRM 客户模型中的应用. 硕士论文. 北京科技大学，2004.

[110] Huber P J. Robust Statistics[M]. New York: Wiley, 1981.

[111] 阎志勇，李明，倪劲锋等. 贝叶斯网络在自适应教育超媒体中的应用[J]. 计算机工程与应用，2002，(8)：217-220.

[112] Breese J, Koller D. Tutorial on Bayesian network [A]. The 14th National Conf on Artificial Intelligence AAAI'97[C]. Rhode Island, 1997.

[113] Murphy K P. A Brief Introduction to Graphical Models and Bayesian Networks. http://www.ai.mit.edu/～murphyk, 2002.

[114] 王飞. 基于遗传算法的 Bayesian 网中连续变量离散化[J]. 计算机学报，2002，25(8)：795-800.

[115] Sung H H, Sung M B, Sang C P. Customer's Time-variant Purchase Behavior and Corresponding Marketing Strategies an Online Retailers Case [J]. Computers & Industrial Engineering, 2002, 43: 801-820.

[116] Boyen X, Friedman N, Koller D. Discovering the Hidden Structure of Complex Dynamic Systems [A]. In: Proceedings of the 15th Cof on Uncertainty in Artificial Intelligence[C]. Stockholm,

Sewden, 1999, 91-100.

[117] Ghahramani Z. Learning Dynamic Bayesian Networks [A]. In: Processing of Sequences and Data Structures[C]. Berlin, Germany, 1998, 168-197.

[118] Binford T, Levitt T, Mann W. Bayesian inference in modelbased machine vision. Proceedings of the 3rd Annual Conference on Uncertainty in Artificial Intelligence, pp. 73-96, 1987.

[119] http://www. cs. huji. ac. il/~talyam/inference. html.

[120] http://www. cs. ubc. ca/labs/beta/Projects/SLS4MPE/.

[121] http://www. mbfys. ru. nl/~jorism/libDAI/.

[122] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.

[123] Rehg J, Pavlovic V, Huang T, Freeman W. Guest editors' introduction to the special section on graphical models in computer vision. Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 785-786, 2003.

[124] Jordan M I. Graphical models. Statistical Science (Special Issue on Bayesian Statistics), vol. 19, pp. 140-155, 2004.

[125] Geman S and Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. Pattern Analysis and Machine Intelligence, vol. 6, no. 6, pp. 721-741, 1984.

[126] Li S. Markov Random Field Modeling in Image Analysis. Springer- Verlag, 2001.

[127] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. International Conference on Machine Learning, pp. 282-289, 2001.

[128] Kumar S and Hebert M. Discriminative random fields. International Journal of Computer Vision, vol. 68, no. 2, pp. 179-201, 2006.

[129] Frey B, Kschischang F, Loeliger H, Wiberg N. Factor graphs and algorithms. Proceedings of the 35th Allerton Conference, 1998.

[130] Rabiner L R. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[131] Smyth P, Heckerman D, Jordan M. Probabilistic independence networks for hidden markov probability models. Neural Computation, vol. 9, pp. 227-269, 1997.

[132] Jensen F V. Bayesian Networks and Decision Graphs. Springer-Verlag, New york, 2001.

[133] Dean T and Kanazawa K. A model for reasoning about persistence and causation. Computational Intelligence, vol. 5, no. 3, pp. 142 - 150, 1989.

[134] Murphy K P. Dynamic bayesian networks: Representation, inference and learning. UC Berkeley, Computer Science Division, PhD thesis, 2002.

[135] Lauritzen S L and Wermuth N. Graphical models for associations between variables, some of which are qualitative and some quantitative. Annals of Statistics, vol. 17, no. 1, pp. 31-57, 1989.

[136] Frydenberg M. The chain graph markov property. Scandinavian Journal of Statistics, vol. 17, pp. 333-353, 1990.

[137] Besag J. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society. Series B (Methodological), vol. 48, no. 3, pp. 259-302, 1986.

[138] Derin H and Cole W S. Segmentation of textured images using using Gibbs random fields. Computer Vision, Graphics and Image Processing, vol. 35, pp. 72-98, 1986.

[139] Murray D and Buxton B. Scene segmentation from visual motion using global optimization. Pattern Analysis and Machine Intelligence, vol. 8, pp. 220-228, 1987.

[140] Hammersley J M and Clifford P. Markov field on finite graphs and lattices. 1971, Unpublished.

[141] Ng A Y and Jordan M. On discriminative vs. generative classifiers: A comparison of logistic

regression and Naive Bayes. Advances in Neural Information Processing Systems, 2002.

[142] Wiberg N, Loeliger H A, Ktter R. Codes and iterative decoding on general graphs. Eur. Trans. Telecomm, pp. 513-525, 1995.

[143] Andersson S, Madigan D, Perlman M. On the markov equivalence of chain graphs, undirected graphs and acyclic digraphs. Scandinavian Journal of Statistics, vol. 24, pp. 81-102, 1997.

[144] Cover T and Thomas J. Elements of Information Theory. John Wiley & Sons, 1991.

[145] Pearl J and Verma T. A theory of inferred causation. Principles of Knowledge Representation and Reasoning, 1991.

[146] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. MIT Press, 2000.

[147] Rissanen J. Hypothesis selection and testing by the MDL principle. The Computer Journal, vol. 42, no. 4, pp. 260-269, 1999.

[148] Schwarz G. Estimating the dimension of a model. Annals of Statistics, vol. 6, no. 2, pp. 461-464, 1978.

[149] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, vol. Series B, 39, no. 1, pp. 1-38, 1977.

[150] Friedman N. The Bayesian structral EM algorithm. Uncertainty in Artificial Intelligence, 1998.

[151] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995.

[152] Yedidia J S, Freeman W T, Weiss Y. Generalized belief propagation. Neural Information Processing Systems, pp. 689-695, 2000.

[153] Shachter R D, Andersen S K, Szolovits P. Global conditioning for probabilistic inference in belief networks. Uncertainty in Artificial Intelligence, pp. 514-522, 1994.

[154] Jensen F V, Lauritzen S L, Olesen K G. Bayesian updating in causal probabilistic networks by local computations. Computational Statistics Quaterly, vol. 4, pp. 269-282, 1990.

[155] Jordan M I, Ghahramani Z, Jaakkola T S, Saul L K. An introduction to variational methods for graphical models. 1999.

[156] Wainwright M and Jordan M. A variational principle for graphical models. In chapter 11 in: New Directions in Statistical Signal Processing. MIT Press, 2005.

[157] Leisink M A R and Kappen H J. General lower bounds based on computer generated higher order expansions. Uncertainty in Artificial Intelligence, pp. 293-300, 2002.

[158] Wainwright M J and Jordan M I. Semidefinite relaxations for approximate inference on grpahs with cycles. Advances in Neural Information Processing Systems, vol. 16, 2004.

[159] Srebro N. Maximum likelihood bounded tree-width markov networks. Uncertainty in Artificial Intelligence, 2001.

[160] Chickering D M, Meek C, Heckerman D. Large-sample learning of Bayesian networks is hard. Uncertainty in Artificial Intelligence, 2003.

[161] Dasgupta S. Learning polytrees. Uncertainty in Artificial Intelligence, 1999.

[162] Narasimhan M and Bilmes J. PAC-learning bounded treewidth graphical models. Uncertainty in Artificial Intelligence, 2004.

[163] Abbeel P, Koller D, Ng A Y. Learning factor graphs in polynomial time and sample complexity. Journal of Machine Learning Research, vol. 7, pp. 1743-1788, 2006.

[164] Bouman C and Liu B. Multiple resolution segmentation of textured images. Pattern Analysis and Machine Intelligence, pp. 99-113, 1991.

[165] Melas D and Wilson S. Double markov random fields and bayesian image segmentation. IEEE Trans. Signal Processing, vol. 50, pp. 357-365, 2002.

[166] Pieczynski W and Tebbache A. Pairwise markov random fields and segmentation of textured images.

Machine Graphics and Vision 9(3), pp. 705-718, 2000.

[167] D'Elia C, Poggi G, Scarpa G. A tree-structured markov random field model for bayesian image segmentation. IEEE Trans. Image Processing, vol. 12, no. 10, pp. 1259-1273, 2003.

[168] Kato Z, Berthod M, Zerubia J. Multiscale markov random field models for parallel image classification. International Conferences on Computer Vision, pp. 253-257, 1993.

[169] Cheng H and Bouman C. Multiscale bayesian segmentation using a trainable context model. IEEE Trans. Image Processing, vol. 10, no. 4, pp. 511-525, 2001.

[170] Wilson R and Li C. A class of discrete multiresolution random fields and its application to image segmentation. Pattern Analysis and Machine Intelligence, vol. 25, no. 1, pp. 42-56, 2003.

[171] Bouman C and Shapiro M. A multiscale random field model for bayesian image segmentation. IEEE Trans. Image Processing, vol. 3, no. 2, pp. 162-177, 1994.

[172] Todorovic S and Nechyba M. Dynamic trees for unsupervised segmentation and matching of image regions. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1762-1777, 2005.

[173] Irving W, Fieguth P, Willsky A. An overlapping tree approach to multiscale stochastic modeling and estimation. IEEE Trans. Image Processing, vol. 6, no. 11, pp. 1517-1529, 1997.

[174] He X, Zemel R, Carreira Perpinan M. Multiscale conditional random fields for image labeling. Computer Vision and Pattern Recognition, vol. 2, pp. II-695- II-702, 2004.

[175] He X, Zemel R S, Ray D. Learning and incorporating top-down cues in image segmentation. European Conference on Computer Vision, pp. 338-351, 2006.

[176] Ren X, Fowlkes C C, Malik J. Cue integration in figure ground labeling. Advances in Neural Information Processing Systems 18, 2005.

[177] Awasthi P, Gagrani A, Ravindran B. Image modeling using tree structured conditional random fields. International Joint Conferences on Artificial Intelligence, 2007.

[178] Reynolds J and Murphy K. Figure-ground segmentation using a hierarchical conditional random field. Proceedings of the Fourth Canadian Conference on Computer and Robot Vision, 2007.

[179] Toyoda T and Hasegawa O. Random field model for integration of local information and global information. Pattern Analysis and Machine Intelligence, vol. 30, pp. 1483-1489, 2008.

[180] Kapoor A and Winn J. Located hidden random fields: Learning discriminative parts for object detection. European Conference on Computer Vision, pp. 302-315, 2006.

[181] Mortensen E N and Jia J. Real-time semi-automatic segmentation using a bayesian network. Computer Vision and Pattern Recognition, pp. 1007 - 1014, 2006.

[182] Alvarado P, Berner A, Akyol S. Combination of high-level cues in unsupervised single image segmentation using Bayesian belief networks. Proceedings of the International Conference on Imaging Science, Systems, and Technology, vol. 2, pp. 675-681, 2002.

[183] Zhang L and Ji Q. Integration of multiple contextual information for image segmentation using a bayesian network. IEEE International Workshop on Semantic Learning Applications in Multimedia, in conjunction with CVPR 2008, 2008.

[184] Feng X, Williams C, Felderhof S. Combining belief networks and neural networks for scene segmentation. Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 467-483, 2002.

[185] Todorovic S and Nechyba M. Interpretation of complex scenes using dynamic tree-structure bayesian networks. Computer Vision and Image Understanding, vol. 106, no. 1, pp. 71-84, 2007.

[186] Freund Y and Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. In Computational Learning Theory: Eurocolt, pp. 23-37, 1995.

[187] Gupta M D, Rajaram S, Petrovic N, Huang T S. Restoration and recognition in a loop. Computer Vision and Pattern Recognition, 2005.

[188] Kumar M P, Torr P H S, Zisserman A. Obj cut. Computer Vision and Pattern Recognition, pp. 18-25, 2005.

[189] Zhang Y and Ji Q. Facial expression recognition with dynamic bayesian networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.

[190] Tong Y,Liao W, Ji Q. Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.

[191] Burl M,Weber M, Perona P. A probabilistic approach to object recognition using local photometry and global geometry. European Conference on Computer Vision, pp. 628-641, 1998.

[192] Weber M, Welling M, Perona P. Unsupervised learning of models for recognition. European Conference on Computer Vision, pp. 18-32, 2000.

[193] Fergus R, Perona P, Zisserman A. Object class recognition by unsupervised scale-invariant learning. Computer Vision and Pattern Recognition, pp. 264-271, 2003.

[194] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, pp. 59-80, 2007.

[195] Arora H, Loeff N, Sorokin A, Forsyth D. Efficient unsupervised learning for localization and detection in object categories. Neural Information Processing Systems, 2005.

[196] Quattoni A, Collins M, Darrell T. Conditional random fields for object recognition. Neural Information Processing Systems, 2004.

[197] Winn J and Shotton J. The layout consistent random field for recognizing and segmenting partially occluded objects. Computer Vision and Pattern Recognition, vol. 1, pp. 37 - 44, 2006.

[198] Hoiem D,Rother C, Winn J. 3d layoutcrf for multi-view object class recognition and segmentation. Computer Vision and Pattern Recognition, 2007.

[199] Kumar S and Hebert M. A hierarchical field framework for unified context-based classification. International Conference on Computer Vision, pp. 1284-1291, 2005.

[200] Li F F and Perona P. A bayesian hierarchical model for learning natural scene categories. Computer Vision and Pattern Recognition, pp. 524-531, 2005.

[201] Quelhas P, Monay F, Odobez J, Gatica-Perez D, Tuytelaars T, Gool L V. Modeling scenes with local descriptors and latent aspects. International Conference on Computer Vision, 2005.

[202] Bosch A, Zisserman A, Munoz X. Scene classification using a hybrid generative/discriminative approach. Pattern Analysis and Machine Intelligence, vol. 30, 2008.

[203] Sudderth E, Torralba A, Freeman W, Willsky A. Describing visual scenes using transformed dirichlet processes. Advances in Neural Information Processing Systems, vol. 19, 2006.

[204] Kivinen J, Sudderth E, Jordan M. Learning multiscale representations of natural scenes using dirichlet processes. International Conference on Computer Vision, pp. 1-8, 2007.

[205] Wang Y and Gong S. Conditional random field for natural scene categorization. British Machine Vision Conference, 2007.

[206] Jain V,Singhal A, Luo J. Selective hidden random fields: Exploiting domain specific saliency for event classification. Computer Vision and Pattern Recognition, vol. 1, 2008.

[207] Gunawardana A, Mahajan M, Acero A, Platt J. Hidden conditional random fields for phone classification. Eurospeech, 2005.

[208] Wang S,Quattoni A, Morency L, Demirdjian D, Darrell T. Hidden conditional random fields for gesture recognition. Computer Vision and Pattern Recognition, 2006.

[209] Torralba A,Murphy K P, Freeman W T, Rubin M A. Context-based vision system for place and object recognition. International Conference on Computer Vision, 2003.

[210] Singhal A, Luo J, Zhu W. Probabilistic spatial context models for scene content understanding. Computer Vision and Pattern Recognition, vol. 1, 2003.

[211] Li L J and Li F F. What, where and who? classifying event by scene and object recognition. International Conference on Computer Vision, 2007.

[212] Murphy K, Torralba A, Freeman W. Using the forest to see the trees: a graphical model relating features, objects and scenes. Neural Information Processing Systems, vol. 15, 2003.

[213] Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S. Objects in context. International Conference on Computer Vision, pp. 1-8, 2007.

[214] Tu Z. Auto-context and its application to high-level vision tasks. Computer Vision and Pattern Recognition, 2008.

[215] Russell B C, Torralba A, Liu C, Fergus R, Freeman W T. Object recognition by scene alignment. Advances in Neural Information Processing Systems, 2007.

[216] Hoiem D, Efros A, Hebert M. Geometric context from a single image. International Conference on Computer Vision, 2005.

[217] Boutell M and Luo J. Incorporating temporal context with content for classifying image collections. International Conference on Pattern Recognition, 2004.

[218] Yu J and Luo J. Leveraging probabilistic season and location context models for scene understanding. ACM conference on Content-based image and video retrieval, pp. 169-178, 2008.

[219] Cao L, Luo J, Kautz H, Huang T. Annotating collections of photos using hierarchical event and scene models. Computer Vision and Pattern Recognition, 2008.

[220] Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T. How flickr helps us make sense of the world: Context and content in community-contributed media collections. ACM Conference on Multimedia, 2007.

[221] Quack T, Leibe B, Gool L V. World-scale mining of objects and events from community photo collections. ACM Conference on Image and Video Retrieval, pp. 47-56, 2008.

[222] Yilmaz A, Javed O, Shah M. Object tracking: A survey. ACM Computing Surveys, vol. 38, no. 4, 2006.

[223] Ross D, Lim J, Lin R S, Yang M H. Incremental learning for robust visual tracking. International Journal of Computer Vision, 2007.

[224] Pavlovic V, Rehg J M, Cham T J, Murphy K P. A dynamic bayesian network approach to figure tracking using learned dynamic models. International Conference on Computer Vision, pp. 94-101, 1999.

[225] Pavlovic V, Rehg J M, MacCormick J. Learning switching linear models of human motion. Neural Information Processing Systems, 2000.

[226] Sigal L, Bhatia S, Roth S, Black M, Isard M. Tracking loose-limbed people. Computer Vision and Pattern Recognition, 2004.

[227] Felzenszwalb P and Huttenlocher D. Pictorial structures for object recognition. Computer Vision and Pattern Recognition, 2000.

[228] Ioffe S and Forsyth D. Probabilistic methods for finding people. International Conferences on Computer Vision, vol. 43, pp. 45-68, 2001.

[229] Mikolajczyk K, Schmid C, Zisserman A. Human detection based on a probabilistic assembly of robust part detectors. European Conference on Computer Vision, 2004.

[230] Mori G, Ren X, Efros A, Malik J. Recovering human body configurations: Combining segmentation and recognition. Computer Vision and Pattern Recognition, 2004.

[231] Ren X, Berg A, Malik J. Recovering human body configurations using pairwise constraints between

parts. International Conferences on Computer Vision, 2005.

[232] Zhang J, Luo J, Collins R, Liu Y. Body localization in still images using hierarchical models and hybrid search. Computer Vision and Pattern Recognition, 2006.

[233] Sudderth E, Ihler A, Freeman W, Willsky A. Nonparametric belief propagation. Computer Vision and Pattern Recognition, 2003.

[234] Zhang L, Chen J, Zeng Z, Ji Q. 2d and 3d upper body tracking with one framework. International Conference on Pattern Recognition, 2008.

[235] Sminchisescu C, Kanaujia A, Li Z, Metaxas D. Discriminative density propagation for 3D human motion estimation. Computer Vision and Pattern Recognition, 2005.

[236] Kim M and Pavlovic V. Conditional state-space model for discriminative motion estimation. International Conference on Computer Vision, 2007.

[237] Ross D, Osindero S, Zemel R. Combining discriminative features to infer complex trajectories. International Conference on Machine Learning, 2006.

[238] Taycher L, Demirdjian D, Darrell T, Shakhnarovich G. Conditional Random People: Tracking humans with CRFs and grid filters. Computer Vision and Pattern Recognition, 2006.

[239] Mccallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation. International Conference on Machine Learning, 2000.

[240] Jordan M and Jacobs R. Hierarchical mixtures of experts and the EM algorithm. Neural Computation, vol. 6, pp. 181-214, 1994.

[241] Kim M and Pavlovic V. Discriminative learning of dynamical systems for motion tracking. Computer Vision and Pattern Recognition, 2007.

[242] Khan Z, Balch T, Dellaert F. An MCMC-based particle filter for tracking multiple interacting targets. European Conference on Computer Vision, 2004.

[243] Rasmussen C and Hager G. Joint probabilistic techniques for tracking multi-part objects. Computer Vision and Pattern Recognition, 1998.

[244] Yu T, Wu Y, Krahnstoever N O, Tu P H. Distributed data association and filtering for multiple target tracking. Computer Vision and Pattern Recognition, 2008.

[245] Su C and Huang L. Spatio-temporal graphical-model-based multiple facial feature tracking. EURASIP Journal on Applied Signal Processing, vol. 2005, no. 13, pp. 2091-2100, 2005.

[246] Hongeng S, Bremond F, Nevatia R. Representation and optimal recognition of human activities. Computer Vision and Pattern Recognition, 2000.

[247] Filipovych R and Ribeiro E. Recognizing primitive interactions by exploring actor-object states. Computer Vision and Pattern Recognition, 2008.

[248] Wang X, Ma X, Grimson W. Unsupervised activity perception by hierarchical bayesian models. Computer Vision and Pattern Recognition, 2007.

[249] Wang X, Tieu K, Grimson W. Correspondence-free multicamera activity analysis and scene modeling. Computer Vision and Pattern Recognition, 2008.

[250] Niebles J, Wang H, Li F F. Unsupervised learning of human action categories using spatio-temporal words. International Journal of Computer Vision, 2008.

[251] Niebles J and Fei-Fei L. A hierarchical model of shape and appearance for human action classification. Computer Vision and Pattern Recognition, 2008.

[252] Zhang D, Perez D, McCowan I. Semi-supervised adapted hmms for unusual event detection. Computer Vision and Pattern Recognition, 2005.

[253] Ivanov Y and Bobick A. Recognition of visual activities and interactions by stochastic parsing. Pattern Analysis and Machine Intelligence, 2000.

[254] Xiang T and Song S. Video behavior profiling for anomly detection. Pattern Analysis and Machine Intelligence, 2008.

[255] Oliver N, Rosario B, Pentland A. A bayesian computer vision system for modeling human interations. Pattern Analysis and Machine Intelligence, 2000.

[256] Xiang T and Song S. Beyond tracking: Modelling activity and understanding behavior. International Journal of Computer Vision, 2006.

[257] Oliver N, Horvitz E, Garg A. Layered representation for human activity recognition. Computer Vision and Image Understanding, 2004.

[258] Duong T, Bui H, Phung D. Activity recognition and abnormality detection with the switching hidden semi-markov model. Computer Vision and Pattern Recognition, 2005.

[259] Lv F and Nevatia R. Single view human action recognition using key pose matching and viterbi path searching. Computer Vision and Pattern Recognition, 2007.

[260] Shi Y and Bobick A. Representation and recognition of activity using propagation nets. International Conferences on Computer Vision, 2003.

[261] Wu J, Osuntogun A, Choudhury T, Philipose M, Regh J. A scalable approach to activity recognition based on object use. International Conferences on Computer Vision, 2007.

[262] Laxton B, Lim J, Kriegman D. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. Computer Vision and Pattern Recognition, 2007.

[263] Vail D, Veloso M, Lafferty J. Conditional random fields for activity recognition. Proceedings of the 2007 Conference on Autonomous Agents and Multiagent Systems, 2007.

[264] Wang L and Suter D. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. Computer Vision and Pattern Recognition, 2007.

[265] Shi Q and Wang L. Discriminative human action segmentation and recognition using semi-markov model. Computer Vision and Pattern Recognition, 2008.

[266] Truyen T, Phung D, Bui H. Adaboost. mrf: Boosted markov random forests and application to multilevel activity recognition. Computer Vision and Pattern Recognition, 2006.

[267] Huang R, Pavlovic V, Metaxas D. A graphical model framework for coupling mrfs and deformable models. Computer Vision and Pattern Recognition, vol. 2, pp. II-739- II-746, 2004.

[268] Huang R, Pavlovic V, Metaxas D N. A hybrid framework for image segmentation using probabilistic integration of heterogeneous constraints. Computer Vision for Biomedical Image Applications, pp. 82-92, 2005.

[269] Tsechpenakis G and Metaxas D. Crf-driven implicit deformable model. Computer Vision and Pattern Recognition, 2007.

[270] Lee C H, Schmidt M, Murtha A, Bistritz A, Sander J, Greiner R. Segmenting brain tumors with conditional random fields and support vector machines. Computer Vision for Biomedical Image Applications: Current Techniques and Future Trends, 2005.

[271] Tu Z, Narr C, Dollar P, Dinov I, Thompson P, Toga A. Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Trans. on Medical Imaging, 2007.

[272] Dietterich T G. Training conditional random fields via gradient tree boosting. International Conference on Machine Learning, 2004.

[273] Torralba A, Murphy K P, Freeman W T. Contextual models for object detection using boosted random fields. Neural Information Processing Systems, 2004.

[274] Sethi A, Rahurkar M, Huang T S. Variable module graphs: A framework for inference and learning in modular vision systems. International Conference on Image Processing, 2005.

[275] Tasker B, Guestrin C, Koller D. Max-margin markov networks. Neural Information Processing

Systems, 2003.
[276] Liu F, Xu D, Yuan C, Kerwin W. Image segmentation based on bayesian network-markov random field model and its application on in vivo plaque composition. IEEE International Symposium on Biomedical Imaging, 2006.
[277] Murino V, Regazzoni C S, Vernazza G. Distributed propagation of a-priori constraints in a bayesian network of markov random fields. Communications, Speech and Vision, IEE Proceedings I, vol. 140, no. 1, pp. 46-55, 1993.
[278] Wang Y, Loe K F, Tan T, Wu J K. Spatiotemporal video segmentation based on graphical models. IEEE Trans. Image Processing, vol. 14, no. 7, pp. 937-947, 2005.
[279] Navaratnam R, Fitzgibbon A, Cipolla R. The joint manifold model for semi-supervised multi-valued regression. International Conference on Computer Vision, 2007.
[280] Urtasun R, Fleet D J, Fua P. 3D people tracking with Gaussian process dynamical models. Computer Vision and Pattern Recognition, 2006.
[281] Lu Z, Carreira M A, Sminchisescu C. People tracking with the Laplacian eigenmaps latent variable model. Neural Information Processing Systems, 2007.
[282] Isard M. Pampas: Real-valued graphical models for computer vision. Computer Vision and Pattern Recognition, 2003.
[283] Ng A and Jordan M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. 2002, in Advances in Neural Information Processing Systems.
[284] Nigam K, McCallum A, Thrun S, Mitchell T. Learning to classify text from labeled and unlabeled documents. in Machine Learning, pages 792-799, 1998.
[285] Jebara T, Kondor R, Howard A. Probability product kernels. Journal of Machine Learning Research, vol. 5, pp. 819-844, 2004.
[286] Holub A, Welling M, Perona P. Combining generative models and fisher kernels for object recognition. International Conference on Computer Vision, pp. 136-143, 2005.
[287] Guo C E, Zhu S, Wu Y N. Modeling visual patterns by integrating descriptive and generative models. International Journal of Computer Vision, vol. 53, pp. 5-29, 2003.
[288] Lester J, Choudhury T, Kern N, Borriello G, Hannaford B. A hybrid discriminative/generative approach for modeling human activities. International Joint Conference on Artificial Intelligence, 2005.
[289] Fritz M, Leibe B, Caputo B, Schiele B. Integrating representative and discriminative models for object category detection. International Conference on Computer Vision, 2005.
[290] Tu Z. An integrated framework for image segmentation and perceptual grouping. International Conference on Computer Vision, 2005.
[291] Lasserre J, Bishop C M, Minka T. Principled hybrids of generative and discriminative models. Computer Vision and Pattern Recognition, 2006.
[292] N'adas A. A decision-theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 31, no. 4, pp. 814-817, 1983.
[293] Jing Y, Pavlovic V, Rehg J M. Efficient discriminative learning of Bayesian network classifier via boosted augmented Naive Bayes. International Conference on Machine Learning, 2005.
[294] Kim M and Pavlovic V. Discriminative learning of mixture of Bayesian network classifiers for sequence classification. Computer Vision and Pattern Recognition, 2006.
[295] de Campos C, Tong Y, Ji Q. Constrained maximum likelihood learning of Bayesian networks for facial action recognition. European Conference on Computer Vision, 2008.
[296] Campos C P D and Cozman F G. Belief updating and learning in semi-qualitative probabilistic

networks. Conference on Uncertainty in Artificial Intelligence, 2005.

[297] Murphy K. Software for graphical models: A review. The ISBA (Interantional Society for Bayesian Analysis) Bulletin, vol. 14, no. 4, pp. 13-15, 2007.

[298] Zhang L,Ji Q. Image Segmentation with a Unified Graphical Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, Issue 8, Vol. 32, pages 1406-1425, 2010.

[299] Koller D,Friedman N. Probabilititic Graphical Models: Principles and Techniques. Cambridge: The MIT Press, 2010.

[300] Pearl J. Probabilistic reasoning in intelligent systems. Morgan Kaufmann,1988.

[301] Cooper G F, Yoo C. Causal discovery from a mixture of experimental and observational data. Proceedings of Uncertainty in Artificial Intelligence, 1999.

[302] Buntine W. Theory refinement on Bayesian Networks. Proceedings of Uncertainty in Artificial Intelligence, 1991.

[303] Friedman N, Koller D. Being Bayesian about network structure. Proceedings of Uncertainty in Artificial Intelligence, 2000.

[304] Lauritzen S. L, Spiegelhalter D. J. Local computations with probabilities on graphical structures and their application to expert systems. J. Royal Statistical Society, 1988.

[305] Darroch J. N and Ratcliff D. Generalized iterative scaling for log-linear models, The Annals of Mathematical Statistics 43 (1972), no. 5, 1470-1480.

[306] Besag J. Efficiency of pseudolikelihood estimation for simple Gaussian fields, Biometrika 64 (1977), 616-618.

[307] Kumar S,Hebert M. Approximate parameter learning in discriminative fields, Snowbird Learning Workshop, 2004.

[308] Sutton C,McCallum A. Piecewise training for undirected models, Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05) (Arlington, Virginia), AUAI Press, 2005, p. 568.

[309] Wainwright M. J, Jaakkola T. S,Willsky A. S. Tree-reweighted belief propagation algorithms and approximate ml estimation 154 by pseudo-moment matching, Workshop on Artificial Intelligence and Statistics, 2002.

[310] Li S. Z. Markov random field modeling in image analysis, Springer,2001.

[311] Welling M. Learning in markov random fields with contrastive free energies, Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 2005, pp. 397-404.

[312] Anguelov D, Taskar B, Chatalbashev V, Koller D, Gupta D, Heitz G, Ng A. Discriminative learning of markov random fields for segmentation of 3d scan data, 2005, pp. II: 169-176.

[313] Narasimhan M. andBilmes J. PAC-learning bounded treewidth graphical models, Uncertainty in Artificial Intelligence, 2004.

[314] Abbeel P, Koller D, Ng A. Learning factor graphs in polynomial time and sample complexity, Journal of Machine Learning Research, vol. 7, pp. 1743-1788,2006.